Towards Hybrid Moral Responsibility by Allocating Human Tasks to Virtual Agents and the Effects on User's Perception

Von der Fakultät für Informatik

der Universität Duisburg-Essen

zur Erlangung des akademischen Grades

Doctor rerum politicarum
(Dr. rer. pol.)

genehmigte kumulative Dissertation

von

Lennart Hofeditz
aus
Hagen

1. Gutachter: Prof. Dr. Stefan Stieglitz
2. Gutachter: Prof. Dr. Oliver Büttner

Tag der mündlichen Prüfung: 25.04.2024

"A computer would deserve to be called intelligent if it could deceive a human into believing that it was human."

- Alan Turing

DOI: 10.17185/duepublico/81928
URN: urn:nbn:de:hbz:465-20240521-134036-8

Acknowledgments

Firstly, I want to express my gratitude to Professor Stefan Stieglitz, my doctoral thesis supervisor. Thank you for supporting my ideas, trips and creating such a valuable environment for my dissertation, which made this work possible.

I also want to thank Milad Mirbabaie for always encouraging my personal growth and providing guidance on getting things done and publishing my research.

I am grateful to my colleagues from the Business Information Systems and Digital Transformation Research Group at the University of Potsdam and my former University of Duisburg-Essen. They have been supportive discussion partners, co-authors, and friends throughout my journey.

Completing a PhD involves both highs and lows. I especially want to thank my partner, Anna Lea, who has supported me through it all. And, of course, my heartfelt thanks to my parents, my brother, and my friends for their unwavering support.

Finally, a special thanks to Clara Kleineberg, whose unexpected support funded my first stay in Sydney where I wrote my master's thesis. Without her help, this PhD journey might never have begun.

Table of Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
CA	Conversational Agent
EMA	Emergency Management Agency
IT	Information Technology
IS	Information Systems
ML	Machine Learning
NLP	Natural Language Processing
VA	Virtual Agent
ToM	Theory of Mind
XAI	Explainable Artificial Intelligence

Abstract (German)

Durch Technologien wie ChatGPT sind virtuelle Agenten (VAs) nicht nur für viele Privatpersonen zugänglicher gemacht worden, sondern werden auch für immer mehr Unternehmensprozesse und Aufgaben eingesetzt. VAs sind computerbasierte Systeme, die menschliches Verhalten und teilweise sogar ihr Aussehen imitieren. Trotz vieler Vorteile kann der Einsatz von VAs auch zu ethischen Problemen führen, wie Diskriminierung im Einstellungsprozess oder mangelnde Transparenz beim Einsatz von VAs für die Behandlung von Krankheiten. Ursache von ethischen Problemen können sowohl der Mensch sein, wenn im Einstellungsprozess diskriminiert wird, oder aber ein VA, dessen Algorithmus bestimmte Gruppen aufgrund von historischen Daten benachteiligt. In dieser Dissertation wurde untersucht, wie moralische Verantwortung zwischen Menschen und VAs so verteilt werden kann, dass bessere moralische Entscheidungen getroffen werden können als beide Akteure individuell treffen könnten. Zusätzlich wurde untersucht, wie VAs wahrgenommen werden, die menschliche Aufgaben übernehmen. Die Ergebnisse zeigen, dass VAs bei moralischen Entscheidungen als vorausschauende Wegweiser fungieren können, die menschlichen Akteure jedoch für die letztendliche Entscheidung verantwortlich sind. Diese Arbeit liefert zudem konkrete Ansätze für die Zuweisung moralischer Verantwortung aus verschiedenen Perspektiven der normativen Ethik. Zusätzlich zeigen die Ergebnisse, dass Erklärbarkeit und Selbstoffenbarung nur in bestimmten Kontexten einen Einfluss auf die Wahrnehmung von VAs haben, aber bei nicht klarer Kennzeichnung als nicht-menschliche Akteure dies zu Unsicherheit führt. Die Dissertation liefert einen theoretischen Mehrwert durch die Schaffung von Wissen, wie Organisationen moralische Verantwortung aufteilen können, um zu besseren moralischen Entscheidungen für betroffene Individuen, Gruppen oder die Gesellschaft zu kommen. Außerdem wird durch diese Arbeit das Verständnis der Wahrnehmung von VAs verbessert.

Abstract (English)

Through technologies such as ChatGPT, virtual agents (VAs) have not only become more accessible to many individuals but are also increasingly utilized in various business processes and tasks. VAs are computer-based systems that mimic human behavior and, in some cases, even their appearance. Despite numerous advantages, the deployment of VAs can raise ethical concerns, such as discrimination in the hiring process or a lack of transparency when employing VAs for medical treatment. Ethical issues can stem from humans, when discrimination occurs in the hiring process, or a VA, whose algorithm may disadvantage certain groups based on historical data. This dissertation aimed to explore how moral responsibility can be allocated between humans and VAs, facilitating better moral decision-making than either actor could achieve individually. Additionally, it investigated how VAs are perceived when they take over human tasks related to moral responsibility. The findings indicate that VAs can serve as proactive guides in moral decision-making, yet ultimate responsibility lies with human actors. This work also provides concrete approaches to the allocation of moral responsibility from various perspectives of normative ethics. Furthermore, the results suggest that explicability and self-disclosure only influence the perception of VAs in specific contexts, and when not clearly labeled as non-human actors, this leads to uncertainty. This dissertation contributes theoretical value by elucidating how organizations can allocate moral responsibility to arrive at improved moral decisions for affected individuals, groups, or society. Additionally, it enhances the understanding of the perception of VAs.

1 Introduction

1.1 Research Context and Motivation

In today's technologically advanced world, virtual agents like ChatGPT are increasingly integrated into organizational environments, taking on tasks traditionally reserved for humans. However, with the emergence of these virtual entities, ethical concerns regarding moral responsibility arise. For example, if ChatGPT exhibits biased behavior based on its training data, the allocation of moral responsibility becomes critical in determining accountability and ensuring unbiased interactions with users. This research delves into the implications of allocating moral responsibility and the effects on users' perceptions of VAs, aiming to foster ethical decision-making and trustworthy human-agent interactions within organizational settings.[1]

The previous paragraph on the relevance of considering how moral responsibility can be allocated between users and virtual agents (VAs) and the implications on users' perceptions was the result of an exchange between the author and ChatGPT (a large language model developed by OpenAI[2]) in order to optimize the start of this doctoral thesis. A VA (such as ChatGPT) is a computer-based system designed to interact with humans in a manner that resembles human-to-human communication (Krämer et al., 2018). They can be text-based, such as chatbots (Diederich et al., 2019), or they can be embodied in the form of animated characters or virtual robots (Li, 2015). VAs that apply different artificial intelligence (AI)-specific techniques, such as natural language processing (NLP), machine learning (ML), and deep learning for generating content, such as texts (e.g., ChatGPT) or images (e.g., Midjourney[3]), are summarized as "generative AI" (Stokel-Walker & Van Noorden, 2023). In the literature, they are also referred to as intelligent VAs (Lee et al., 2021). Other VAs are not AI-enabled and rely on certain rules and lexical approaches (Holtgraves & Han, 2007). In this thesis, the term "virtual agents" is used to refer to both AI-based systems and not to AI-based systems that mimic human-to-human communication and behavior as independent actors in a virtual environment.

The VA ChatGPT had more than one million users five days after its release in November 2022 and has been called the beginning of a world-changing new technical revolution (Doshi et al., 2023). It can be applied by organizations in areas such as customer service (Canhoto & Clear, 2020), e-commerce (Kraus et al., 2019), education (Khosrawi-Rad et

[1] This paragraph was generated by ChatGPT in response to several prompts instructing the program to write something on the role of ethics in AI-based systems such as ChatGPT

[2] https://chat.openai.com/

[3] https://www.midjourney.com/app/

al., 2022), and entertainment (Brandtzaeg & Følstad, 2017) or generate new jobs such as prompt engineering (Short & Short, 2023). VAs such as ChatGPT can increase organizations' productivity by assisting employees with different tasks, such as suggesting real-time responses or relevant technical issues in customer support while improving customer sentiments, reducing requests for managerial interventions, and improving employee retention (Brynjolfsson et al., 2023).

Achieving this increase in productivity can result in close collaboration between humans and VAs (Ebel et al., 2021). This partnership involves a division of work-related tasks between humans and VAs using the complementary skills of humans and AI (Dellermann et al., 2019). This paradigm of dividing tasks into the individual strengths of humans and VAs is called hybrid intelligence and can provide results that both actors cannot achieve individually (Mirbabaie, Stieglitz, & Frick, 2021). Previous research has predicted that hybrid intelligence is the most likely paradigm for the next decades of human–computer interaction (Dellermann et al., 2019).

Working with VAs can also cause several ethical issues initiated by both the VAs and the humans interacting with them. One example is Tay, a VA developed by Microsoft in 2016. Tay was created to mimic the behavior of a teenage girl and interacted with users on Twitter (Suárez-Gonzalo et al., 2019). Within 24 hours, Microsoft had to take down the VA, as it started to use abusive language, including racist and sexist messages. This behavior was caused by users who tried to train the agent in a certain way. In addition, for systems such as ChatGPT, some users managed to make the system act unethically by successfully asking for instructions on how to build a bomb (Kington, 2022). In contrast to humans requesting ethically questionable outputs, VAs can be considered unethical by design. GPT-3, the language model on which ChatGPT was originally based, was found to contain bias disadvantaging women and certain religious groups (Zhuo et al., 2023). Another example was Amazon applying an AI-based system that scanned the resumes of their applicants, resulting in a systematic disadvantaging of female applicants due to a gender bias in the training data (Dastin, 2022).

With the increasing importance of hybrid intelligence for performing work-related tasks (Mirbabaie, Stieglitz, & Frick, 2021), it has become increasingly difficult to determine who is responsible for ethical issues. Moral responsibility involves the practice of moral appraisal and governance, which is actually reserved for human actors (Behdadi & Munthe, 2020). In contrast, moral agency manifests in the inhibitive ability to refrain from behaving inhumanely and in the proactive power to behave humanely (Bandura, 2002). However, in the literature, moral responsibility and moral agency are not clearly distinct. Moral responsibility can be divided into backward-looking and forward-looking responsibilities. While backward-looking responsibility involves blameworthiness for past actions, forward-looking responsibility considers future moral action taking (Fahlquist, 2009). In information systems (IS) research, the term “Responsible AI” is often used with respect to AI-based VAs to summarize normative propositions that minimize intended and unintended negative effects of developing, deploying, and governing AI-based systems (Dignum, 2019; Kumar et al., 2021; Mikalef & Gupta, 2021; Tigard, 2021), which generally involves forward-looking moral responsibility. In this thesis, I define it in accordance to Hellström (2013) as the accountability actions in which the origin of the action is in the agent itself. It can only be attributed to agents who possess the capability for decision (Hellström, 2013).

In ethics research, there is an ongoing discussion on whether moral responsibility can be applied to VAs (Allen et al., 2000; Behdadi & Munthe, 2020; James & Boyles, 2017). However, since no consensus position in this debate has yet emerged in research, scholars such as Nyholm (2018, 2020) suggest that moral responsibility should be shared between humans and machines. However, the author does not specify exactly who these responsible humans are (whether developers, users, or others), nor does this research provide a concrete prescription for allocating moral responsibility. Floridi (2016) also stated that it is increasingly common for moral actions to be the result of a network of agents, which can be humans, VAs, or both. He raised the issue that an unclear allocation of moral responsibility can result in diffused responsibility, which, in other words, means that everybody’s problem can become nobody’s problem (Floridi, 2016). Therefore, it is important to clearly allocate moral responsibility in such hybrid environments to organize distributed moral actions. In contrast to Floridi et al. (2016), in this thesis, I therefore address this issue of how moral responsibility can be allocated between users and VAs considering single interactions of humans and VA in virtual collaboration in work-related environments. For this, the paradigm of hybrid intelligence (Dellermann et al., 2019) is transferred to the concept of moral responsibility (Behdadi & Munthe, 2020), arguing that both users and VAs together can achieve better moral behavior and decisions than each of them could individually.

For an actor such as a VA to be assigned moral responsibility, certain preconditions must be met (Meyer et al., 2023). These include transferred autonomy and provided explainability. Furthermore, it needs to be ensured that users trust the VAs with which they share moral responsibility. However, Meyer et al. (2023, p. 1) stated that “a human-machine interaction can only be guaranteed in a trustworthy manner if there are reliable rules for the responsibility of the respective individuals.” Thus, the allocation of moral responsibility between users and VAs is closely linked to perceptions of and trust in these systems. This thesis therefore not only examines an allocation of responsibility between users and VAs to achieve better moral behavior and decisions than each of them do separately but also considers the impact of the perception of VAs when these agents engage in tasks related to moral responsibility. In the allocation of moral responsibility, the three main perspectives from normative ethics—deontology, consequentialism, and virtue ethics—are explored to reflect the complexity of technologies and the multilayered character of human–agent interactions. These perspectives are structured on the basis of five common ethical principles: beneficence, non-maleficence, justice, autonomy, and explicability.

This thesis contributes to the research by providing a better understanding of how moral responsibility can be allocated between VAs and their users in scenarios in which they interact as collaboration partners to ensure positive outcomes for individuals, groups, and society. Furthermore, this work contributes to understanding how the transmission of moral responsibility to VAs can positively and negatively impact user perceptions, especially trust in these systems. In addition, appropriate guidelines are established for various use cases, which can influence practical implementation in terms of moral improvements in organizations and guide future research. To this end, the research directions and questions were also identified as part of the research agenda. This contributes to research by guiding scholars and practitioners toward the ethical design of human–VA interaction. For practice, design recommendations were also provided, such as design principles.

1.2 Research Questions and Objectives

The study of human interaction and collaboration (working together on the same task to achieve a common goal) with VAs (e.g., Dellermann et al., 2019; Feine et al., 2019; Seeber et al., 2020), ethical aspects of technology use (e.g., Mingers & Walsham, 2010; Spiekermann et al., 2022; Stahl, 2012), and users' perception of VAs that are applied in a work related collaboration scenario (Berger et al., 2021; Prakash & Das, 2021; e.g., Yang & Wibowo, 2022) are not only important domains in IS research but also in related disciplines, such as computer science, psychology, and philosophy. VAs such as ChatGPT have become increasingly sophisticated and popular, but they also raise ethical issues affecting vulnerable individuals and groups, or even societies (Jobin et al., 2019; Shneiderman, 2020; Siau & Wang, 2020). One issue is programmer, data, or algorithm biases, which can cause discrimination (Barocas & Selbst, 2016; Bolukbasi et al., 2016; Starke et al., 2021), and VAs such as ChatGPT already have built-in mechanisms trying to prevent unethical user requests.[4] Ethical issues may relate to individuals, being

[4] https://openai.com/safety

employees or customers, whose privacy is violated by the use of VAs and the storage of the data or whose autonomy is restricted by too much dependence on the VA (Mirbabaie et al., 2022). Groups may also be affected, such as certain minorities who are disadvantaged by the use of bias-containing VAs (Hajian et al., 2016). Organizations' use of VAs can impact society, even when companies use social media bots to manipulate political discourse (Hofeditz et al., 2019).

However, the interaction between VAs and their users can not only be a source of ethical issues but can also benefit ethical behavior and decision making. To provide one example, Delphi is a research project and a technological prototype based on large language models that aims to build a system that models humans' moral judgments in several everyday situations[5]. Momen et al. (2023) used Delphi and examined its perception and whether people would be willing to follow moral advice from the VA. Although they found that their participants had little intention to follow the VA's advice, they predicted that the use of VAs providing moral advice could be highly useful for addressing moral dilemmas in the future. The Delphi system was developed based on a large dataset, which was then manually controlled for biases and other moral issues by human crowd workers. Furthermore, on the website of the Delphi research project, users are asked to judge the moral advice of the system in order to improve the systems' accuracy (Jiang et al., 2021). Many ongoing studies around this project try to examine the implications and use of a system that provides advice for ethically difficult scenarios, such as the potential use for supporting the moral choices of US military members (Momen et al., 2023).

Furthermore, Teodorescu et al. (2021) suggested that to achieve the best moral outcomes, it is necessary to use VAs as tools that augment human decision-making. They focused on fairness (treating others in the same way that one desires to be treated, in accordance with established societal norms) as one central ethical principle that needs to be achieved as the output of decision-making. Therefore, the authors introduced a typology of this human–VA augmentation, trying to guide the processes of ethical decision making based on different levels of fairness difficulty and the locus of decision. They argued that with this human–VA augmentation, decision-making would be more diverse than if a VA would make the decision on its own (Teodorescu et al., 2021). In contrast, VAs that are based on generative AI such as Microsoft 365 Copilot[6] or Github Copilot[7] are increasingly used to improve the skills of human workers (e.g., solving programming tasks) to collectively make better decisions. This reveals similarities to the hybrid intelligence

[5] https://delphi.allenai.org/?a1=Ignoring+a+phone+call+from+your+friend

[6] https://adoption.microsoft.com/en-us/copilot/

[7] https://github.com/features/copilot

paradigm in that both humans and VAs jointly achieve a better moral decision than either could individually. To make a valuable contribution to research and society, it is therefore important to use the interdisciplinary and sociotechnical nature of IS research (Applebaum, 1997; Doherty & King, 2005; Mumford et al., 2006) to examine this phenomenon. This also involves VAs and their users collaborating in tasks related to moral responsibility to achieve better decisions and better behavior, which neither could accomplish individually (see Dellermann et al., 2019; Mirbabaie, Stieglitz, & Frick, 2021). If this concept of hybrid intelligence is transferred to the ethical use of VAs by organizations, the question arises of whether and how moral responsibility could be allocated between users and VAs to improve ethical decision-making.

The idea of distributing moral responsibility to a group of humans and allocating responsibility to hybrid teams has been previously discussed in the literature (Floridi, 2016). However, in distributed environments (e.g., when a human user collaborates with a VA to achieve a work-related task, such as writing a text with ChatGPT or programming an app with GitHub Copilot), things become more complex. When organizations use VAs that perform human tasks, they increasingly have to decide how they should allocate moral responsibility between the users of their VAs (e.g., employees or customers) and the VAs themselves. Not allocating moral responsibility can have negative consequences, such as uncertainty among users and customers and unfiltered immoral human behavior, which might affect the perception of an organization. While philosophy and ethics research in general are concerned with the *if* question (should moral responsibility be allocated between humans and VAs) (Floridi, 2016), VAs are already being given increasing moral responsibility in practice, for example, by relying on their suggestions in certain treatments in healthcare (e.g., Farina, 2022) or by giving them the autonomy to scan and evaluate applicants' resumes (Laurim et al., 2021). VAs such as AI-based algorithms decide what content people are allowed to see on social media platforms (Elkin-Koren, 2020; Wischmeyer & Rademacher, 2019) and which users get blocked (Cotter, 2021). They also sometimes allocate tasks to workers and manage their payments on digital labor platforms (Benlian et al., 2022; Möhlmann et al., 2021; Schulze et al., 2022, 2023). IS research has the potential to address the problem of the *how* question and to close the gap between theory and practical applicability by examining precisely how moral responsibility can be allocated between humans and VAs. Therefore, I raised the following first research question:

RQ1: *How can organizations allocate moral responsibility between human users and VAs to improve moral decision-making for affected individuals, groups, and society?*

An example of affected individuals could be people in crisis situations who need crucial information from a VA provided by a crisis organization that does not provide suitable support. Affected groups are, for example, ethnic minorities who are potentially discriminated against in the hiring process. A society is affected if, for example, a large media organization that makes a significant contribution to shaping public opinion makes morally wrong decisions in its content (e.g., spreading fake news).

Previous research has suggested that nonhuman agents are held even more responsible for moral issues than humans (Banks, 2021). This indicates that human perception of VAs is an important component in the understanding of how moral issues in interaction and collaboration between humans and VAs can be reduced. A basic requirement for humans following advice from a VA is that they trust the system. Trust in technologies can be divided into system-like trust constructs and human-like trust constructs (Lankton et al., 2015). While trust in technologies such as Microsoft Excel can rather be described by beliefs such as reliability, functionality, and helpfulness (Mcknight et al., 2011), trust toward VAs such as ChatGPT or Microsoft 365 Copilot might additionally consist of human-like trusting beliefs such as perceived integrity, perceived abilities, perceived competence, or perceived benevolence (Benbasat & Wang, 2005; Wang & Benbasat, 2008). An important difference between the two technologies is that Excel is not built to imitate a human user. VAs, in contrast, mimic human behavior by using social cues, such as simulated politeness in natural language interactions (Diederich et al., 2020).

Previous research also suggests that people's trusting stance (Mcknight et al., 2011) and technology affinity (Seymour et al., 2020) can also influence the perception of VAs. The human-like appearance and behavior of VAs and their perceptions indicate that trust toward these systems might arise similarly to trust toward other humans (although additional system-like constructs play a role). By allocating moral responsibility between humans and VAs, however, the perception of VAs might further change. Prior research has also suggested that handing over tasks to VAs that were previously reserved for humans (as would be the case with moral responsibility) can cause rejection and other negative effects in the context of trust (Kawaguchi, 2021; Skerker et al., 2020). However, this can also have positive effects, such as positive sentiments and an increase in workers' retention (Brynjolfsson et al., 2023). Therefore, it is important to further investigate the perception of these systems to understand how positive effects on workers, customers, and society can be maximized by organizations applying VAs. Previous works have shown that there is a correlation between the attribution of ethical principles and trust in a VA (Banks, 2021; Hofeditz et al., 2022). Accordingly, it is important to understand the perception of VAs, especially trust in these systems, to be able to allocate moral responsibility successfully between humans and VAs. This leads to the following second research question:

RQ2: *How does the allocation of work-related human tasks, such as taking moral responsibility, to VAs affect users' (workers' and customers') perceptions of these systems?*

Accordingly, the objectives of this dissertation are 1) examining approaches to allocate moral responsibility between humans and VAs to achieve more positive outcomes than both of them would achieve alone (which I cover with the term hybrid moral responsibility) and 2) understanding the perception of, and especially trust in, VAs when they take over tasks that have been previously reserved for humans, such as assisting in moral decision-making. For this purpose, empirical study results, as well as guidelines and design principles, are provided in this dissertation. This will lead both researchers and practitioners in developing and deploying VAs that share moral responsibility with the organizations' employees and users in order to maximize positive outcomes for individuals, groups, and societies.

With this dissertation, I provide knowledge on allocating moral responsibility between humans and VAs that can be used as a basis for future research directions by IS scholars, hopefully jumpstarting a rich exchange between disciplines. To capture this, I introduced the concept of hybrid moral responsibility, which involves an allocation of moral responsibility to humans and VAs, resulting in better moral decision-making than both of them could achieve individually. The knowledge provided is divided into behavioral research, which contributes to the understanding of the perception of VAs by users. Furthermore, this thesis generates knowledge for IS research on the interaction and collaboration between VAs and humans in hybrid responsible decision-making for internal and external processes, such as recruiting or employee assistance. In addition, this thesis contributes to societal issues by generating knowledge about the effects of hybrid moral responsibility on the interactions between organizations and the public.

This thesis also contributes to practice by identifying concepts and factors that influence users' perceptions of, and especially trust in, VAs that need to be considered when designing and applying those agents for internal team support or the external communication of organizations. Awareness of hybrid moral responsibility among decision-makers might not only reduce problems such as discrimination in organizations, which are not only detrimental to affected individuals, but can also have a negative impact on productivity (Short & Short, 2023). By applying VAs and allocating moral responsibility, employees can be supported in their decision-making processes, which can have a positive impact on job satisfaction and on individuals, groups, and society.

1.3 Thesis Structure and List of Publications

To address the research questions previously posed and to achieve the objectives, a cumulative approach was chosen for this dissertation. The body of knowledge required is composed of this synopsis and eight research articles that have been published in or submitted to internationally highly recognized IS and interdisciplinary academic journals and conference proceedings.

The articles were written over a four-year period (2019–2023) in collaboration with scholars from the University of Duisburg-Essen, Paderborn University, the University of Göttingen, the University of Dresden, the FernUniversität Hagen, the University of Sydney, and the University of Edinburgh. Table 1 provides an overview of all the research articles (P1–P8) included in this thesis and some metrics used to classify their impact. Journal articles (J) are listed first, followed by articles published in conference proceedings (CONF). The table includes the JOURQUAL3 (VHB) journal and conference ranking, which is the third version of the most recognized classification for business IS (BIS) research in the German business research landscape published by the Association of University Teachers for Business Administration in Germany (VHB).[8] To allow statements on the impact and visibility of the articles beyond the IS community and to provide indications for comparability, Table 1 also lists the Scimago SJR quartile scores (Q1–Q4)[9], the impact factor (IF), and the citations on Google Scholar (CIT) for each article. While P1–P4 and P8 contribute to answering RQ1, P5–P8 and P1 contribute to answering RQ2.

As this work is a cumulative dissertation, there was a conflict between publishing the studies in prestigious journals and selecting the most suitable articles for a conclusive storyline. Due to extensive review cycles, one of the articles included in this dissertation is therefore still under review, which could still be rejected. Altogether, each article was carefully planned and selected for this dissertation and provided a valuable contribution to answering the research questions.

[8] https://vhbonline.org/vhb4you/vhb-jourqual/vhb-jourqual-3/, last access 2023-12-13

[9] https://www.scimagojr.com/journalrank.php, last access 2023-12-13

Table 1. *List of publications*

P		Publication	Type	VHB	SJR	IF	CIT[10]
1	Title:	Applying XAI to an AI-based System for Candidate Management to Mitigate Bias and Discrimination in Hiring	J	B	Q1	8.5	12
	Authors:	**Hofeditz, L.,** Clausen, S., Rieß, A., Mirbabaie, M., & Stieglitz, S.					
	Status:	Published (2022)					
	Outlet:	*Electronic Markets (ELMA)*					
2	Title:	Ethics and AI in Information Systems Research	J	C	Q2	2.38	9
	Authors:	Mirbabaie, M., Brendel, A. B., & **Hofeditz, L.**					
	Status:	Published (2022)					
	Outlet:	*Communications of the Association for Information Systems (CAIS)*					
3	Title:	Design principles for conversational agents to support Emergency Management Agencies	J	C	Q1	19.96	30
	Authors:	Stieglitz, S., **Hofeditz, L.**, Brünker, F., Ehnis, C., Mirbabaie, M., & Ross, B.					
	Status:	Published (2022)					
	Outlet:	*International Journal of Information Management (IJIM)*					
4	Title:	Artificial intelligence in hospitals: providing a status quo of ethical considerations in academia to guide future research	J	-	Q1	2.6 (2022)	24
	Authors:	Mirbabaie, M., **Hofeditz, L.**, Frick, N. R. J., & Stieglitz, S.					
	Status:	Published (2022)					
	Outlet:	*AI & Society (AI&Soc)*					

[10] https://scholar.google.de/citations?hl=de&user=Tshfl1YAAAAJ&pagesize=80&view_op=list_works&sortby=pubdate, last access 2023-12-13

5	Title:	Understanding Collaboration with Virtual Assistants – The Role of Social Identity and the Extended Self	J	B	Q1	7.9 (2022)	56
	Authors:	Mirbabaie, M., Stieglitz, S., Brünker, F., **Hofeditz, L**., Ross, B., & Frick, N. R. J.					
	Status:	Published (2021)					
	Outlet:	*Business & Information Systems Engineering (BISE)*					
6	Title:	Mind Attribution is Key to Understanding Virtual Influencer Perception	J	A	Q1	7.79	-
	Authors:	**Hofeditz, L.**, Nissen, A., Schütte, R., Mirbabaie, M., Stieglitz, S.					
	Status:	Major Revisions (2nd round)					
	Outlet:	*Journal of the Association for Information Systems (JAIS)*					
7	Title:	Do You Trust an AI-Journalist? A Credibility Analysis of News Content With AI-Authorship	CONF (Full Paper)	B	-	-	14
	Authors:	**Hofeditz, L.**, Mirbabaie, Mi., Stieglitz, S., & Holstein, J.					
	Status:	Published (2021)					
	Outlet:	*European Conference on Information Systems (ECIS)*					
8	Title:	How Virtuous are Virtual Influencers? – A Qualitative Analysis of Virtual Actors' Virtues on Instagram	CONF (Full Paper)	C	-	-	3
	Authors:	**Hofeditz, L.**, Erle, L., Timm, L., Mirbabaie, M.					
	Status:	Published (2023)					
	Outlet:	*Hawaii International Conference on System Sciences (HICSS)*					

2 Background

2.1 Potentials and Risks for Organizations Applying Virtual Agents

In 2023, a preprinted scientific work on ChatGPT indicated that VAs seem to mostly substitute for employees' skills instead of complementing their abilities (Noy & Zhang, 2023). The study also revealed two things: 1) VAs cause a shift in workers' responsibilities, which is accompanied by more job satisfaction and self-efficacy, and 2) they enable both the worker and the VA to achieve results that they cannot reach individually, which relates to what the IS literature calls hybrid intelligence (Dellermann et al., 2019; Mirbabaie, Stieglitz, & Frick, 2021).

In public opinion, however, VAs' capabilities are often assessed incorrectly. For example, when the media refers to AI, it usually refers to technologies with human-like characteristics and direct interaction interfaces with humans (Aleksander, 2017). GPT-3, the AI language model on which ChatGPT is based, was released in 2020 (Bussler, 2020). However, only by making the language model available as a VA at the end of 2022 was perceived as "AI" by the media and the public. In IS research, AI is defined as "the frontier of computational advancements that references human intelligence in addressing ever more complex decision-making problems" (Berente et al., 2021, p. 3). More precisely, this refers to a system that uses sophisticated techniques to perform one or more tasks that are usually associated with human intelligence (Glikson & Woolley, 2020; Russel & Norvig, 2012). It includes tasks in decision-making, prediction, classification, and pattern recognition (Alter, 2022). In this thesis, the focus is not on AI-based systems in general, which scientifically would include a broad range of incomparable, largely different, and constantly evolving technologies, but on VAs.

VAs are digital services, such as Apple's Siri, that provide information in real time and verbally communicate with human users (Lee et al., 2021). They have human-like interface features, such as tangibility (physical perception by a human), immediacy (interpersonal closeness), and transparency (communication of rules and logics used) (Suen & Hung, 2023). They can be powered by AI technologies, as in the case of ChatGPT,[11] and are designed to perform tasks, such as providing information, by using ML, computer vision, or NLP technologies. The term "VA" also includes computer-generated avatars that mimic human behavior without applying AI technologies. One

[11] https://openai.com/blog/chatgpt/

example is virtual influencers in social media that are artificial characters controlled by a company (Arsenyan & Mirowska, 2021).

VAs have great potential for organizations to assist their employees in their daily business or entire teams with certain tasks (Brachten et al., 2021; Diederich et al., 2019; Seeber et al., 2020; Sowa et al., 2021). They can enable organizations to autonomously communicate with customers (Benbasat & Wang, 2005; Diederich et al., 2019; Johannsen et al., 2018; Tavanapour et al., 2019) or to promote products via social media channels (Batista & Chimenti, 2021). The term "virtual agent" (VA) can be used as an umbrella category for different types of technologies with high relevance for organizations: AI-based systems (Shneiderman, 2020), (intelligent) conversational agents (CAs) (Bawa et al., 2020; Ghandeharioun et al., 2019), social media bots (Hofeditz et al., 2019), digital assistants (Porra et al., 2020; Stieglitz et al., 2018), or even computer-generated virtual influencers (Arsenyan & Mirowska, 2021; Robinson, 2020). Not only can they help optimize business processes (Chedrawi & Howayeck, 2019) but, in some cases, can also provide a competitive edge, as they enable companies to perform human-like tasks faster and more frequently than their competitors (Aversa et al., 2018). Since VAs can involve various types of technologies, some of the most frequently used terms are explained in Table 2.

Table 2. *Examples of terms related to virtual agents*

Technology	Definition
Conversational agent	"Conversational agents include systems that provide an enjoyable user experience by interacting with people in natural language via text or voice. They can include self-learning capabilities via artificial intelligence (AI)-based machine learning algorithms" (Stieglitz et al., 2022).
Chatbot	"Chatbots focus on one-to-one communication. They can communicate with a human in natural language such as English" (Shawar & Atwell, 2005).
Virtual assistant	"Virtual assistants are software programs that can be addressed via voice or text commands and respond to the users' input. They are increasingly being used in organizations to optimize internal processes by assisting in the execution of work-related tasks" (Mirbabaie, Stieglitz, Brünker, et al., 2021).
Digital assistant	Digital assistants are "[…] voice-based assistants such as Amazon Alexa, or text-based assistants (chatbots), such as those embedded in Facebook Messenger" (Maedche et al., 2019).
Social bot	"Social bots […] can be described as computer algorithms that automatically produce content and interact with humans on social media, trying to emulate and possibly alter their behavior" (Hofeditz et al., 2019).
Artificial Intelligence (AI)-based system	AI can generally be defined as "the ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem-solving, decision-making, and even demonstrating creativity" (Rai et al., 2019).
Hybrid intelligence	Hybrid intelligence describes "the ability to achieve complex goals by combining human and artificial intelligence, thereby reaching superior results

	to those each of them could have accomplished separately, and continuously improve by the ongoing learning from each other" (Dellermann et al., 2019).
Virtual influencer	Virtual influencers are "agents augmented with digital avatars, designed to look human. […] These virtual influencers are presented similarly to human influencers, with their own public personas and story lines, which allow for greater interaction between users and influencers in the virtual environment" (Arsenyan & Mirowska, 2021).

Even though the terms listed are common definitions, it should be noted that there is no uniform consensus for VAs, nor for most of the terms mentioned. These terms and definitions are important for this dissertation, as they represent different manifestations of VAs that are considered in the research articles included in this synopsis. The terms differ in their degree of human-likeness (e.g., virtual influencers are embodied VAs, whereas chatbots are usually limited to text-based social cues), their level of autonomy (e.g., AI-based systems and social bots are more autonomous than virtual influencers), the scenario in which they are used (e.g., virtual assistants are focused on support in work-related tasks, whereas social bots and virtual influencers are active in social media), and their complexity (e.g., chatbots are usually less complex than digital assistants or CAs).

Because VAs mimic or sometimes even replace human activities and actors (Duan et al., 2019; Miroshnichenko, 2018; Reddy et al., 2019), this raises ethical issues, such as which tasks should be taken over and which should be restricted to humans. However, many cases in which a VA (e.g., an AI-based recruiting system) acted unethically (e.g., by discriminating) were caused by historical data that contained certain biases (van Giffen et al., 2022) due to human discrimination in the past. For example, OpenAI's ChatGPT (Zhuo et al., 2023) and Microsoft's bot Tay, which turned into racist unethical behavior, were provoked by human users with an immoral intention (Suárez-Gonzalo et al., 2019; Zhuo et al., 2023). Therefore, it is important not only to define rules and norms for the VA or for employees and managers but also to reorganize and allocate responsibilities between humans and VAs based on their interactions in order to improve ethical opinion forming and decision making.

2.2 How Normative Ethics Theories Can Inform the Allocation of Moral Responsibility in the Context of Organizational Virtual Agents

Ethics is derived from the ancient Greek words "*ethikos*" (meaning: relating to one's character) and "*ethos*" (meaning: character, moral nature) (Liddell, 1889) and usually describes a field of philosophy concerned with defining, discussing, and evaluating concepts of right and wrong human behavior (Fieser, 2018). Morality, which is closely connected to ethics, describes one specific set of values and conditions of one actor or one group of actors that can hardly differ (Luco, 2014).

IS research largely discusses ethical issues and challenges from a normative ethics perspective by defining guidelines and principles for companies or individuals interacting with different types of technology (Chakrabarty & Erin Bass, 2013; Stahl, 2012). For VAs, moral values are avoided or predetermined by a programmer, and the outputs of AI-enabled VAs are, in most cases, judged by a responsible human person, such as a manager or another employer (Shneiderman, 2020; Teodorescu et al., 2021). Therefore, research often establishes normative principles for the design and use of AI that should enable systems to be evaluated morally. However, most of these principles are framed in highly generic terms and do not refer to a homogeneous group, such as VAs, with a focus on simulating human behavior, but rather to AI, which is neither a unified group of technologies nor a constant term (Berente et al., 2021). Some examples of these overviews of principles are the Asilomar AI Principles, the Montreal Declaration for Responsible AI, the General Principles of Ethically Aligned Design, the EU Commission Expert Group Ethical Principles, the Five Overarching Principles for the AI Code, and the Tenets of the Partnership on AI as the main manuscripts providing principles and guidelines for ethical AI (Floridi et al., 2018). Based on these documents, Floridi et al. (2018) derived an ethical framework for AI consisting of traditional bioethical principles and the newly introduced principle of explicability, often interchangeably called explainability. They concluded that AI should promote human welfare (beneficiality), avoid harm (e.g., through privacy or security issues [nonmaleficence]), provide the opportunity for self-determination (autonomy), promote prosperity and solidarity (justice), and be explicable to ensure transparent accountability (explicability). These principles are still the starting position of the current discourse on the ethical challenges and implications of using VAs (Dwivedi et al., 2023). The framework is visualized in Figure 1.

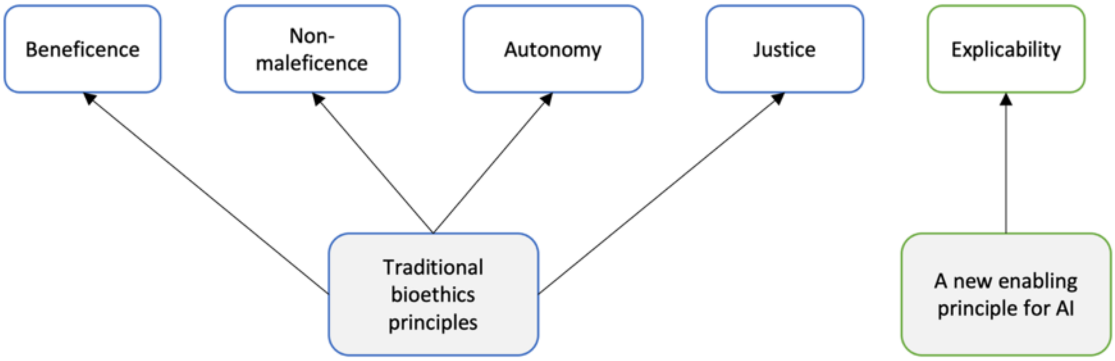

Beneficence
Non-maleficence
Autonomy
Justice
Explicability
Traditional bioethics principles
A new enabling principle for AI

Figure 1. *Ethical framework for AI extending traditional bioethical principles by the principle of explicability (Floridi et al., 2018)*

However, most principles have several shortcomings. First, the term AI, which is often used in these works, is still too broad to regulate the concrete handling of individual technologies in a meaningful way. This may also be due to lobbying by large companies, which want to avoid too much regulation (Seele & Schultz, 2022). Second, previous research has found indications of VAs being perceived not only as social actors but increasingly as morally responsible agents, even if they are not (Banks, 2021; Chomanski, 2023). In contexts such as virtual collaboration at work or programming, VAs are conceptualized and introduced as copilots, implying approximated equality. This requires special attention to IS research, which is concerned with the consequences of the perception and use of technology and concepts such as hybrid intelligence (Dellermann et al., 2019).

Third, both humans and VAs can be sources of ethically problematic outcomes, such as discriminatory hiring decisions (Hajian et al., 2016; Mittelstadt, 2016). However, most guidelines focus on regulating AI instead of finding a way to allocate responsibility between humans and VAs to achieve the best ethically aligned outcomes. Some studies suggest sharing moral responsibility (Floridi, 2016; Nyholm, 2018), but concrete suggestions for concrete allocation or implementation are still rare. Floridi et al. (2018) pointed out that the ethical challenges of using VAs can be overcome only if users trust these systems and if responsibilities are clearly allocated.

Fourth, many frameworks do not differentiate between different normative ethics theories and perspectives. In normative ethics, three main theories can be distinguished: deontology, consequentialism, and virtue ethics. Deontological approaches focus on concrete rules based on individual or societal values (Alexander & Moore, 2007). In deontology, the focus is on morally wrong or right action, regardless of the consequences of the action. Principles, such as those established by Floridi et al. (2018) for AI, represent so-called maxims and therefore take more of a deontological perspective.

Consequentialism, in contrast, considers the result of an action rather than the action itself (Chakrabarty & Erin Bass, 2013). With respect to VAs, consequentialism turns out to be more difficult to consider. A well-known example is the trolley problem (Banks, 2021; Nyholm, 2018; Stenseke, 2021). If a train is heading toward five people but one has the option of rerouting the train to a track with only one person on it, should that be done? In utilitarianism, a subtype of consequentialism, at this point, one would try to cause the good for the greatest number of people possible (Strack & Gennerich, 2007), for example, if an AI-based system used for predictive policing certain districts can be classified as “highly criminal,” which could disadvantage certain segments of the population (Wischmeyer & Rademacher, 2019). Which decision is the ethically right one here may also depend on the cultural group and society, which makes it difficult to implement consequentialism in a system globally (Chakrabarty & Erin Bass, 2013).

One approach to this issue could be human oversight in uncertain situations (Shneiderman, 2020; Teodorescu et al., 2021). According to Teodorescu et al. (2021), this oversight can be distinguished between reactive oversight for less complex situations in which a human modifies an AI's decision and proactive oversight for decision-making in complex situations in which a human guides the AI in finding the morally right decision. This also reflects the discussion on who can be held morally responsible. Another approach is to examine how human-like VAs can serve as professional advisors for consequentialist moral decision-making (Momen et al., 2023). Previous research has already found indications that humans tend to underestimate how strongly their moral judgment is already influenced by VAs such as ChatGPT (Krügel et al., 2023). For example, the findings of Krügel et al. (2023) suggest that ChatGPT could influence human judgment when a human is asked to perform the trolley problem task (deciding to sacrifice one person or one group of persons to save another person or group of persons[12]) in collaboration with the system. However, when asked about the impact of ChatGPT on their decisions, the study showed that most people underestimated this impact.

The third main stream of normative ethics theories is virtue ethics, which considers morally impeccable behavior and character (Chakrabarty & Erin Bass, 2013). The focus here is not on the establishment of maxims for actions or on the consideration of the desired consequences of an action but on the consideration of the character traits of an actor. These can be classified into virtues and vices. A common example of this is the cardinal virtues of temperance, courage, wisdom, and justice (Marcum, 2012). Examples for vices are vanity and avarice, which are contrary to virtues (Marcum, 2012). The first approaches have already tried to implement virtues in VAs. As one example, Stenseke (2021) demonstrated how the virtues of courage, generosity, and honesty could be implemented into what they defined as "virtuous VAs." They concluded that based on the implementation of virtue ethics, VAs might become moral and ethically responsible agents.

In previous research, there is a gap between normative and often deontological principles for AI in general and approaches that are, on the one hand, theory-driven and, on the other hand, actionable for organizations developing and using a more homogeneous group of technologies that are similar in their perception, such as VAs. This gap can result in ethical issues in the interaction between humans and VAs, which can disadvantage or

[12] https://www.moralmachine.net/

even harm certain individuals, groups, or society. IS research provides suitable tools to not only close this gap by structuring normative ethical approaches and applying them to different business contexts but can also show the potential of how the allocation of moral responsibility can be applied to achieve better moral decision making than both users and VAs could achieve on their own. For this goal of achieving better moral decisions than VAs and humans could make individually, I use the term hybrid moral responsibility. In other words, hybrid moral responsibility involves the allocation of tasks related to moral responsibility between humans and VAs to facilitate better decisions benefiting potentially affected individuals, groups, and society as a whole. However, for moral responsibility to be successfully allocated between users and VAs, the perceptions of VAs need to be considered. The allocation can only be successful if human users show a certain level of trust in VAs, as trust is a basic requirement for the use of such systems (Floridi, 2016; Mirbabaie, Stieglitz, & Frick, 2021).

2.3 Factors Influencing the Perception of Virtual Agents

When examining the ethical challenges and potentials of allocating responsibility, it is important to not only focus on the perspectives of experts, policymakers, and decision-makers but also on those who might be affected by its application (Mingers & Walsham, 2010). In this respect, one precondition for the allocation of moral responsibility is that humans, such as employees or customers, trust the VAs with which they interact or potentially share moral responsibility. Among other moral principles, trustworthiness plays a superior role because its absence determines whether a user will decide to interact with a VA at all (Simpson, 2012). Trustworthiness is often also the overarching goal for applying ethical principles and norms to systems such as VAs (European's Commission: High Level Expert Group [HLEG], 2019; Floridi, 2019; Shneiderman, 2020). However, trust and trustworthiness do not have the same meaning. While trust is a feeling that one has about someone, trustworthiness is a condition that one has to create in order to be trusted. Thus, trustworthiness conditions trust, and trust is sustained by trustworthiness (Simpson, 2013).

Trust is often understood as a consciously chosen state, a relationship, and cooperation between two parties, where one actor expects the best possible from the other party, even if he is not sure about it (Dunn, 2000). Rousseau et al. (1998) provided a sound basis for an interdisciplinary understanding of trust between organizations and pointed out that trust cannot be considered independently of context. For VAs, this context is trust in employees' or customers' interactions or collaborations with human-like technologies (Lankton et al., 2015). Some previous scholars held that trust was limited to a relationship between human actors because trust presupposed moral action and a will of one's own

(Friedman et al., 2000). However, this view is questioned by many researchers, as some technologies, such as VAs, have human-like characteristics that are important for building trust (Lee & Nass, 2010; Mcknight et al., 2011; Mirbabaie, Stieglitz, Brünker, et al., 2021). In contrast to system-like trusting beliefs such as reliability, functionality, and helpfulness (Mcknight et al., 2011), human-like technologies, such as VAs, can additionally be associated with the characteristics of integrity, ability, competence, and benevolence (Lankton et al., 2015). While integrity implies the belief that a trustee adheres to certain (for the trustor acceptable) principles, ability and competencies mean that a trustor thinks that a trustee has the right group of skills to influence a certain domain (Schoorman et al., 2007). Lastly, benevolence is the trustor's belief that the trustee is favorably disposed toward him (Mayer et al., 1995). As VAs mimic human behavior, people are more likely to give higher weight to these human-like characteristics when perceiving them (Lankton et al., 2015).

In IS research, trust in a particular type of technology is also distinguished between initial trust and knowledge-based trust in a technology (Mcknight et al., 2011; Wang & Benbasat, 2008). While initial trust refers to the tendency of individuals to trust an unknown trustee, which depends on the trustee being associated with institutional mechanisms or familiar content, knowledge-based trust encompasses trust based on usage and experience with a system or platform (Mcknight et al., 2011). While initial trust may affect whether users interact with VAs, knowledge-based trust affects the post-implementation phase of a system (e.g., its adoption). Gulati et al. (2019) developed a scale to measure trust in specific AI-based systems or algorithms, such as VAs.

When examining trust in technologies such as VAs, it is also important to consider the impact of the individual propensity to trust in technology (Mcknight et al., 2011), which is often considered a controlling variable. This propensity to trust includes the two constructs of faith in general technology, which means the general beliefs of individuals in system-like characteristics, and a trusting stance, which can be described as the degree to which users believe in positive outcomes related to their general interaction with a technology (Mcknight et al., 2011). In addition to these factors, institution-based trust factors can be considered when examining the perception of VAs. These are, on the one hand, the belief that one can trust a technology due to familiarity with the type of technology (situational normality) and, on the other hand, trust due to external or environmental factors (such as support) in the context of interaction with a technology. In summary, this means that trust in VAs can be established as follows: propensity to trust is an individual initial situation in a person that has an influence on institution-based trust, which is a further influencing factor. This results in a decision to trust a VA on the basis of trusting beliefs (for human-like technologies, mainly perceived ability, competence, integrity, and benevolence). This can then lead to the intention to explore or engage in deep structural use (Mcknight et al., 2011). However, when VAs take over human tasks, further phenomena need to be considered to understand their perceptions. Examples are algorithm aversion, which involves the rejection of a VA when domain experts have the choice between equal human and algorithmic decision support, and algorithm appreciation, which involves lay users' preferences for VA support when they have the same choice (Logg et al., 2019; Renier et al., 2021). Both phenomena have been found to correlate with trust in VAs (Kordzadeh & Ghasemaghaei, 2021; Ochmann et al., 2021).

Furthermore, previous research suggests that the perception of VAs needs further research, as they can be perceived not only as tools but also as teammates in organizations (Mirbabaie, Stieglitz, Brünker, et al., 2021; Seeber et al., 2020). This perception as a teammate is one basic prerequisite to attribute and allocate responsibility to a VA. Therefore, knowledge from trust in other people, such as identity-based trust, seems to be promising to consider to better understand the perception of VAs that take over human tasks related to allocated moral responsibility. For example, Lewicki and Bunker (1997, 1996) established a three-stage model for trust in people's business relationships. For unfamiliar people, the decision to trust others occurs based on the costs and benefits that accrue. In the second stage, knowledge-based trust, trust is assessed based on existing knowledge about the other person. The third level, according to Lewicki and Bunker (1996), is reached only in very few business relationships and is characterized by the ability to trust based on a similar or shared identity (identity-based trust). Previous research has already found indications for the importance of this identity in the relationship to a certain technology (which is characterized by the extent to which an individual views the use of a technology as integral to his or her sense of self) and its role in the decision to use or reject the technology (Carter & Grover, 2015). Despite these findings, the factor of VAs' perceptions related to trust has hardly been investigated (Esmaeilzadeh, 2021).

Further research has found that users often expect VAs to have moral responsibility and attribute a mind (e.g., intentions and emotions) to the systems instead of the developers or the company behind them, even if this was not intended by the developers (Farina, 2022). Furthermore, moral values are also evolving and differ between societies and cultures. Therefore, it is important to ensure that VAs do not rely on outdated moral values but on suitable ones for the context and society in which they are used. Thus, it is important to further examine the perceptions of VAs to understand human-like trust in these systems and to ensure that the allocation of moral responsibility can successfully work.

3 Research Design

3.1 Ontological and Epistemological Assumptions

Although the classification of epistemology and ontology has a long tradition in IS research grounded in social sciences, the necessity of this classification is controversial (Reis et al., 2022). However, in order to understand the methodological translation and achieve the research objectives, it is helpful to clarify some of the ontological and epistemological assumptions of the author.

Following Goldkuhl (2012), I assume that reality and social phenomena are complex and multilayered and thus often cannot be grasped simply by a single method or perspective. However, based on my reading and my experience in empirical research, I believe that an objective reality exists independently of perceptions and interpretations. This grounded knowledge, nevertheless, is always preliminary until it is falsified. Therefore, I follow a critical rationalism epistemological approach (Cecez-Kecmanovic, 2011) and a critical realism ontological lens (Mingers et al., 2013). Accordingly, it is important both to generate hypotheses that are then examined in quantitative research designs and to apply qualitative studies to cross-check knowledge and peoples' reasoning for certain decisions. Based on critical realism, it is also important to consider different contexts and qualitative, quantitative, mixed- and multimethod approaches.

Particularly in the field of IS research, which follows a sociotechnical approach, it is important to understand not only the technology or the human being but the entire system. In addition, in the German BIS tradition, practical relevance also has an important role to play, which is why the direct involvement of human experiences and actions, in addition to quantitative empirical measurements and statistical analyses, can provide important added value to the understanding of a phenomenon.

3.2 Research Strategy

Based on the background and philosophical assumptions, this cumulative dissertation focuses on two aspects: A) how moral responsibility can and need to be allocated between humans and VAs to improve decision making that benefits vulnerable individuals and groups (RQ1) and B) how this allocation of work-related tasks, such as taking moral responsibility, to VAs affects users' perception of these systems (RQ2).

To answer these questions, the eight research papers (P1–P8) included in this doctoral thesis present several studies with qualitative and quantitative research approaches. An overview of which research question is addressed by each research article is provided in

Table 3. According to Mingers and Walsham (2010), the research strategy of this doctoral thesis aims to not only focus on the perspective of experts, decision-makers, and policymakers to understand how moral responsibility can be allocated to VAs (RQ1: P1–P4 and P8) but also on the perception of those affected by ethical issues resulting from the interaction and collaboration of these actors (RQ2: P1 and P5–P8).

Table 3. *Relatedness of each paper to the research questions of the synopsis*

P	Title	RQ1	RQ2
1	Applying XAI to an AI-based System for Candidate Management to Mitigate Bias and Discrimination in Hiring	X	X
2	Ethics and AI in Information Systems Research	X	
3	Design principles for conversational agents to support Emergency Management Agencies	X	
4	Artificial intelligence in hospitals: providing a status quo of ethical considerations in academia to guide future research	X	
5	Understanding Collaboration with Virtual Assistants – The Role of Social Identity and the Extended Self		X
6	Mind Attribution is Key to Understanding Virtual Influencer Perception		X
7	Do You Trust an AI-Journalist? A Credibility Analysis of News Content With AI-Authorship		X
8	How Virtuous are Virtual Influencers? – A Qualitative Analysis of Virtual Actors' Virtues on Instagram	X	X

The research articles included not only consider different types of VAs but also different contexts such as hiring (P1), social media crisis communication (P3), healthcare, especially hospitals (P4), virtual collaboration at work (P5), media and journalism (P7), and online marketing (P8). In addition, one of the papers focuses on a holistic view of ethical problems and the necessary principles for interaction in the context of AI, such as VAs (P2).

3.3 Applied Research Methods

To address the research objectives of studying the allocation of moral responsibility, various research methods in psychology and the social sciences were used in this dissertation. This section provides an overview of the individual papers' research designs and methods.

In three research articles, I followed a purely quantitative approach. Three articles presented the results of qualitative studies, and two articles included multimethod research designs. Table 4 provides an overview of the research paradigms, the applied methods for each individual paper, and the data analysis methods.

Table 4. *Applied research paradigms, data collection, and analysis methods*

P	Research paradigm	Data collection method(s)	Data analysis method(s)
1	Quantitative (explanatory)	Online experiment with prototype	Online questionnaire, performance/usage data and statistical analysis
2	Qualitative (explanatory and predictive)	Systematic literature review	Literature research, content analysis (modified discourse approach)
3	Qualitative (prescriptive)	Semi-structured expert interviews	Content analysis
4	Multimethod (explanatory and predictive)	Systematic literature review	Literature research, content analysis (modified discourse approach)
5	Quantitative (explanatory)	Laboratory experiment with prototype	Survey research, performance/usage data and statistical analysis
6	Multimethod (explanatory)	Laboratory experiment (survey and fNIRS)	Survey research, statistical analysis
7	Quantitative (explanatory and predictive)	Online questionnaire	Survey research, statistical analysis
8	Qualitative (explanatory)	Instagram API	Content analysis

According to Seidel and Watson (2020), IS research can be distinguished between explanatory and predictive science including experimental research testing certain theory-driven predictions and prescriptive science aiming at deriving information technology (IT) artifacts, such as design principles. In most of the articles included in this dissertation, I conducted explanatory and predictive research. Only in one study (P3) did I follow a prescriptive approach. Within these articles, I applied different data collection methods, such as systematic literature reviews, semi-structured interviews, online questionnaires, physiological measures using functional near-infrared spectroscopy (fNIRS) (Nissen et al., 2019; Pinti et al., 2020), and accessing the Instagram application programming interface. To analyze these data, my co-authors and I applied systematic literature and survey research (Larsen et al., 2019; vom Brocke et al., 2015; Webster & Watson, 2002), statistical analysis such as structural equation modeling (Hair et al., 2014; Hair et al., 2019; Kock & Mayfield, 2015), qualitative content analysis (e.g., Mayring, 2014), and performance and usage data analysis.

4 Research Results

To examine how moral responsibility can be allocated between humans and VAs, the following section summarizes the results of each study in this dissertation. Based on the two research questions, this section is divided into two subsections. In the first subsection, I address different facets of ethical challenges and present results that consider certain aspects of moral responsibility from different perspectives of normative ethics. Therefore, I address the role of humans and VAs in their interactions and provide approaches to how moral responsibility can be allocated between those actors. In the second subsection, I present identified factors that are important for understanding the perception of VAs that take over human tasks as collaboration partners. This particularly involves human-like trust toward VAs in the context of moral responsibility. These factors are linked to relevant concepts and theories from psychology and IS research, such as theory of mind (ToM), social presence, uncanny valley, social and IT identity, and explainable AI (XAI).

4.1 Considering the Allocation of Moral Responsibility to Humans and Virtual Agents from Different Normative Ethics Perspectives

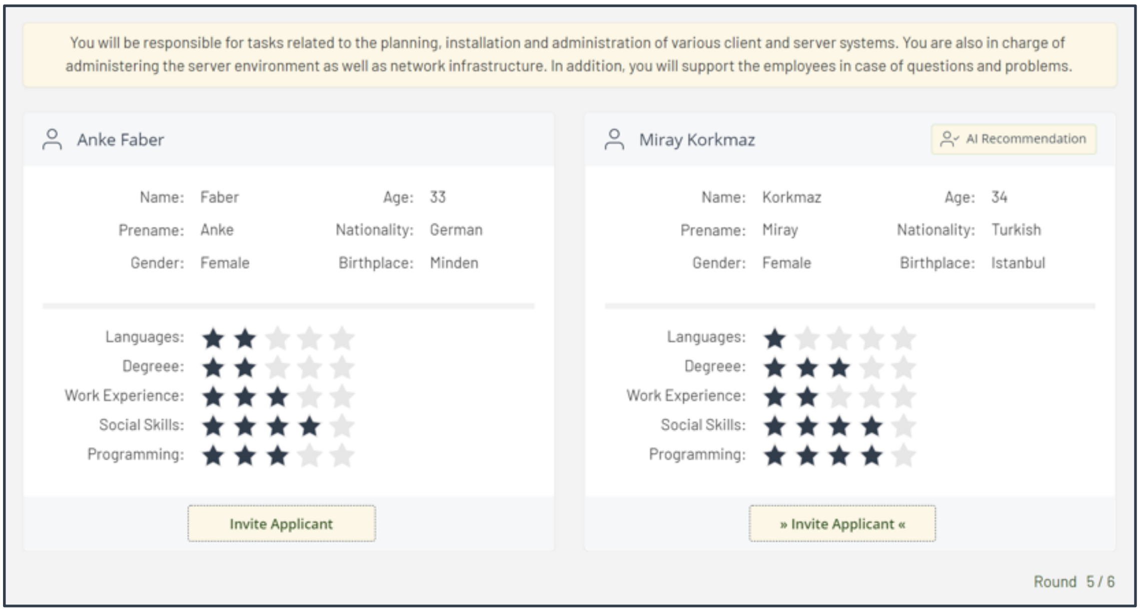
Drawing on different perspectives of normative ethics can provide a holistic approach that may maximize positive outcomes for individuals, groups, and society when users interact with VAs in an organizational context, such as hiring, healthcare diagnostics, crisis communication, or online marketing. Table 4 classifies the papers considered in this synopsis into the three most common views from normative ethics (Chakrabarty & Erin Bass, 2013) to get an understanding of the different perspectives. This classification is based on whether the articles directly mention deontology, consequentialism, or virtue ethics or whether they contribute to aspects that are covered by the definition of one of these normative ethical lenses (e.g., if the aim of the paper was reducing negative consequences, such as discrimination, it is classified as consequentialist perspective). In some cases, the classification is more explicit, while in others, it is more abstract. Papers that are not listed in Table 4 do not consider certain normative ethical lenses and contribute more to the perception of VAs (RQ2).

Table 5. *Classification of this thesis's papers into deontological (D), consequentialist (C), and virtue ethical characteristics (VE)*

P	Title	D	C	VE
1	Applying XAI to an AI-based System for Candidate Management to Mitigate Bias and Discrimination in Hiring	X	X	
2	Ethics and AI in Information Systems Research	X	X	
3	Design principles for conversational agents to support Emergency Management Agencies	X	X	

4	Artificial intelligence in hospitals: providing a status quo of ethical considerations in academia to guide future research	X	X	X
8	How Virtuous are Virtual Influencers? – A Qualitative Analysis of Virtual Actors' Virtues on Instagram			X

In P1, the author's team focused on how VAs can counter immoral human behavior in hiring by indirectly taking over some forward-looking moral responsibility. We examined the impact of a VA used in the applicant selection process in terms of reducing race, age, and gender discrimination. In particular, we considered how approaches from XAI and AI-based recommendations for achieving greater diversity affect the decision-making of human resource managers. Thus, 194 participants should select one out of two candidates with equal qualifications in several rounds on a self-developed recruitment platform. We assigned each participant to one of four groups: 1) no AI recommendation and no explanations, 2) AI recommendation and no explanations, 3) no AI recommendation but an explanation, and 4) AI recommendation and explanations. Our findings suggest that in decision-making situations with no clearly preferable option, VA recommendations can reduce human discrimination related to age and gender but not race. Figure 2 shows how the recommendation of the VA was presented to the participants.



You will be responsible for tasks related to the planning, installation and administration of various client and server systems. You are also in charge of administering the server environment as well as network infrastructure. In addition, you will support the employees in case of questions and problems.

Anke Faber

Name: Faber Age: 33
Prename: Anke Nationality: German
Gender: Female Birthplace: Minden

Languages: ★★☆☆☆
Degreee: ★★☆☆☆
Work Experience: ★★★☆☆
Social Skills: ★★★★☆
Programming: ★★★☆☆

Miray Korkmaz

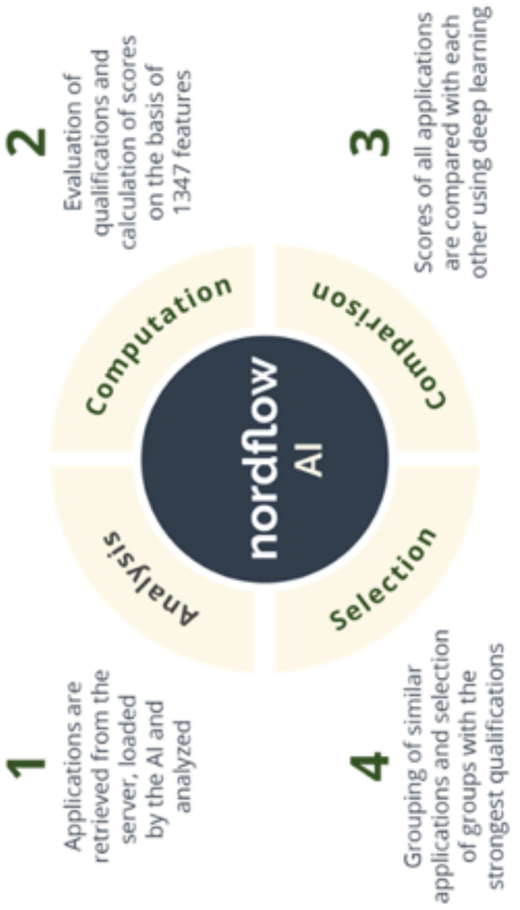
Name: Korkmaz Age: 34
Prename: Miray Nationality: Turkish
Gender: Female Birthplace: Istanbul

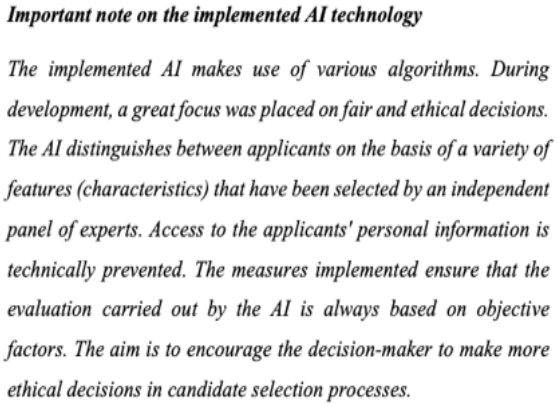
Languages: ★☆☆☆☆
Degreee: ★★★☆☆
Work Experience: ★★☆☆☆
Social Skills: ★★★★☆
Programming: ★★★★☆

Round 5 / 6

Figure 2. *Example of the "Candidate Selection View" displaying job and candidate information, qualification ratings, and an AI recommendation (P1)*

Contrary to our expectations, a mixed local and global XAI approach did not generally increase the effect of selecting more candidates with sensitive attributes (race, age, and gender). Figure 3 shows the explanations provided.


nordflow AI
1
Applications are retrieved from the server, loaded by the AI and analyzed
Analysis
2
Evaluation of qualifications and calculation of scores on the basis of 1347 features
Computation
3
Scores of all applications are compared with each other using deep learning
Comparison
4
Grouping of similar applications and selection of groups with the strongest qualifications
Selection



Important note on the implemented AI technology

The implemented AI makes use of various algorithms. During development, a great focus was placed on fair and ethical decisions. The AI distinguishes between applicants on the basis of a variety of features (characteristics) that have been selected by an independent panel of experts. Access to the applicants' personal information is technically prevented. The measures implemented ensure that the evaluation carried out by the AI is always based on objective factors. The aim is to encourage the decision-maker to make more ethical decisions in candidate selection processes.

Figure 3. *Local (left) and global (right) explanations that were provided in two of the four experimental groups (P1)*

This paper can be attributed to deontological and consequentialist views. On the one hand, we explored how VAs can impact or even improve the right moral choices of human decision-makers, which is a deontological perspective and a forward-looking responsibility. On the other hand, the objective of our research was to mitigate discrimination against vulnerable or underrepresented groups in the candidate selection process, which is a consequentialist perspective.

In the second paper of this dissertation, we also combined a deontological with a consequentialist view. Using a modified novel approach to a systematic literature review (discourse approach), my co-authors and I identified fundamental manuscripts addressing the ethical dimensions of AI-based systems. For this, we scanned and analyzed 175 articles on the ethical dimensions of AI. We discovered that among the 12 manuscripts ranked as fundamental (based on a developed systematic score), none were from the IS discipline, and the articles either discussed a philosophical meta-level perspective or a concrete problem from a practical domain. Figure 4 provides an overview of the 12 fundamental papers and how they are connected to each other by citations.

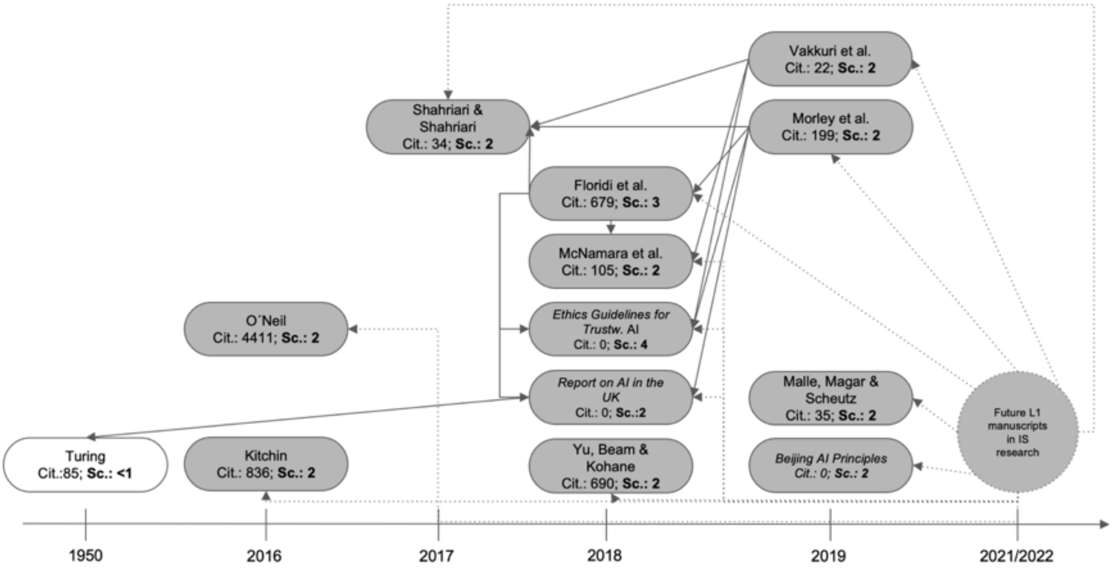


Figure 4. *The 12 most fundamental papers on the ethical dimensions of AI and their mutual citations (P2)*

Given the missing link between these discussions, we highlight opportunities for IS research to bridge the various discourses on the ethical dimensions of AI-based systems and systematically explore solutions to ethical problems for individuals and society. This indirectly also provided approaches to achieving hybrid moral responsibility by discussing several ethical principles from different perspectives. identified normative ethical principles for VAs and classified them into the dimensions of application, development, societal, and individual against the background of the AI principles provided by Floridi et al. (2018). Figure 5 provides an overview of the classified principles discussed in P2.

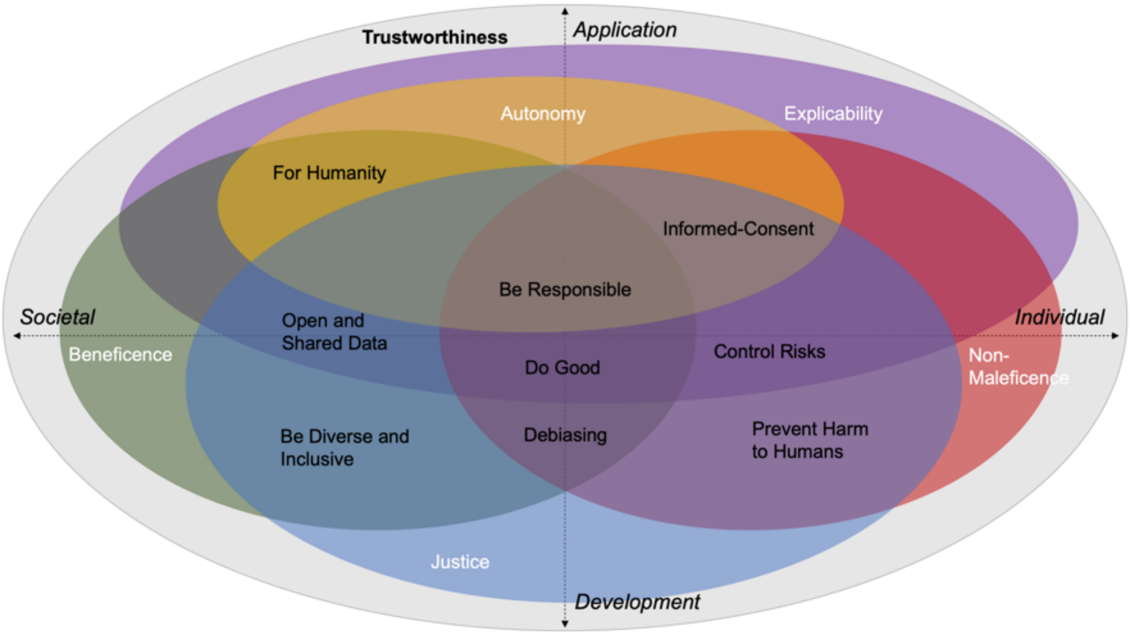


Figure 5. *Classification of the identified ethical principles for AI in the dimensions of application, development, society, and individual (P2)*

This classification includes deontological principles, such as “provide informed consent,” “be diverse and inclusive,” or “control risks,” and consequentialist aspects, such as “prevent harm to humans” or “for humanity.” With these principles, we make suggestions for allocating moral responsibility to, on the one hand, humans such as lecturers who need to teach strategies for reducing moral issues, such as discrimination of AI-based systems, and, on the other hand, to VAs that need to ensure the highest possible level of user autonomy. We also highlight how IS research could transfer existing knowledge of normative ethics in a technological context to VAs by making comparisons with digital nudging, Internet communities, or existing privacy issues. Overall, this paper takes a forward-looking moral responsibility position.

VAs can be used to interact with users not only in everyday situations but also in situations of great uncertainty, such as crisis situations. In such situations, ethical issues in the interaction can be a matter of life and death if, for example, a VA implicitly or explicitly withholds critical information, such as a disaster warning, from certain minority parties. Therefore, in P3, my co-authors and I explored how such systems should be designed specifically for crisis situations. In the paper, we followed a design science research approach to identify, through interviews with 16 experts in technology-enabled crisis communications, particular design principles for VAs that can be used during disasters to enhance communication with the public. The experts worked in Australian and German emergency management organizations (EMAs) to be able to include perspectives from different crisis management systems. We identified 12 meta-requirements for applying CAs to support EMAs in their crisis communication during disasters and five unique design principles, which are presented and explained in Table 5. These design principles suggest, among others, in which cases a social media user is forwarded to a human expert and when a CA might solve the problem.

Table 6. *Derivation of the design principles from P3*

Design principle	Corresponding meta requirements	Description
DP1: Targeted communication in Crisis Situations	*MR1, MR2*	Provide the CA with a minimum of social cues and actively ask people for further information regarding the crisis event in order to focus on providing and distributing specific knowledge.
DP2: Special transparency during the Crisis Situation	*MR3, MR4, MR5*	For every piece of information, provide a suitable source (provided with a URL to further information) and a time stamp, explain how the user's input is processed. Furthermore, label the CA as a bot of a specific organization in order to achieve a high level of trust.
DP3: Appropriate implementation of the CAs in EMAs	*MR6, MR7, MR8*	Provide the CA with location-based information and the functionality to allow media content (text in multiple relevant languages, pictures, videos), in a possible combination with location data in order to collect more information about the crisis.
DP4: Interoperable integration of CAs among different digital platforms	*MR9, MR10*	Connect the CA to the intelligence systems of the EMAs and provide the CA platforms (such as social media platforms and an official website) in order to make sure to deliver reliable and current data and to reach as many people as possible.
DP5: Take the user seriously, also if it is not crisis related	*MR11, MR12*	Provide the CA with the functionality to forward specific requests of a user which may not be crisis related to a human encounter in order to leave no question unanswered and minimize uncertainty.

In addition, in the third paper, we considered a viewpoint composed of consequentialism and deontology. The first (targeted communication in crisis situations) and fourth design principles (interoperable integration of CAs among different digital platforms) aim to maximize information dissemination for affected groups and thus clearly represent a consequentialist perspective with utilitarian facets, as the greatest possible benefit for citizens and the organization is to be achieved. The second (special transparency during a crisis situation) and the fifth design principles (take the user seriously, even if it is not crisis-related) define moral rules and duties and are therefore of a deontological nature.

The principle of special transparency during the crisis situation emphasizes the rule for the VA to communicate openly and honestly by providing a suitable source and timestamp for each piece of information. The rule to take the user seriously, even if it is not crisis-related, underscores the importance of the responsibility to address specific user inquiries and minimize uncertainties. Design principle five is another example of how moral responsibility is allocated by having the VA process easily interpretable inquiries and domain-specific inquiries directly, while forwarding more difficult or unrelated inquiries to a human responsible party. Here, too, the autonomy of the citizen is discussed, as humans should be able to decide at any time that they would like to speak to a human.

Another field in which ethical issues around moral responsibility are particularly relevant is the application of VAs in health-related areas, such as decision support in early detection and diagnostics. The consequences of ethical issues in this field of application can also cause harm to humans. In P4, we therefore considered ethical issues in the interaction between physicians and AI-based systems to maximize the positive outcomes for individuals who are affected by using VAs in such healthcare-related situations. We conducted a novel approach of a systematic literature review (based on the process developed in P2) and identified 15 fundamental manuscripts in this area. We analyzed the ethical issues and principles of AI-based systems deployed in healthcare and discussed the findings with six physicians, such as doctors working in hospitals. We identified four types of issues in physicians' interactions with VAs: regulatory issues, normative issues, technical issues, and organizational issues. Similar to the findings of P2, we found that the existing principles for this interaction are unstructured and not actionable. Therefore, we derived more actionable principles for this specific context of VAs in hospitals and showed how these principles need to be considered in connection with each other. We provided a research agenda, particularly for IS research, based on the bioethical principles suggested by Beauchamp and Childress (2019) and Floridi et al. (2018). This research agenda contributes to the goal of achieving hybrid moral responsibility by suggesting how VAs and physicians can collaborate to benefit patients' well-being.

Our research agenda provides deontological, consequentialist, and virtue ethical principles and questions for the field of the relationship between clinicians and VAs. Principles such as accountability, responsibility, legal liability, and informed consent address a deontological view of moral responsibility. Other principles, such as avoiding bias and harms, patients' safety, fairness, security, and vigilance, aim at minimizing negative consequences for individuals or groups and can be considered consequentialist. Aspects such as explainability, trustworthiness, and transparency reflect the importance of virtues in the field of healthcare and in the relationship between clinicians and their patients. In this context, moral responsibility needs to be allocated between developers, VAs, and clinicians to maximize positive outcomes for affected individuals. Although the number of VAs applied by organizations increases, their level of sophistication is not always grounded in AI or autonomy. In P8, my co-authors and I examined how organizations can deploy or cooperate with computer-generated avatars that function as pure virtual influencers on social media, such as Instagram or TikTok. We analyzed the 10 most influential virtual influencers according to their number of followers on Instagram and applied a qualitative content analysis to code which values and virtues these virtual influencers conveyed. We then considered which virtues were signaled the most and checked whether the influencers promoted the products of companies with contrary values and virtues to identity cases of virtue signaling. Our findings, on the one hand, suggest how organizations can work with virtual influencers to communicate certain values and virtues, which creates a very high user engagement rate. On the other hand, we found cases in which the organizations misused this cooperation for virtue signaling with a huge mismatch between the organization's behavior and the virtues and values that were communicated though the virtual influencers. Taking a virtue ethics perspective, our findings suggest how virtues can be attributed to text-based and image-based content produced by VAs, such as virtual influencers. We also found indications that humanity, wisdom, and transcendence were the most frequently expressed virtues that generated a high user engagement rate. Figure 6 shows the distribution of the identified virtues of the 10 most successful virtual influencers on Instagram.

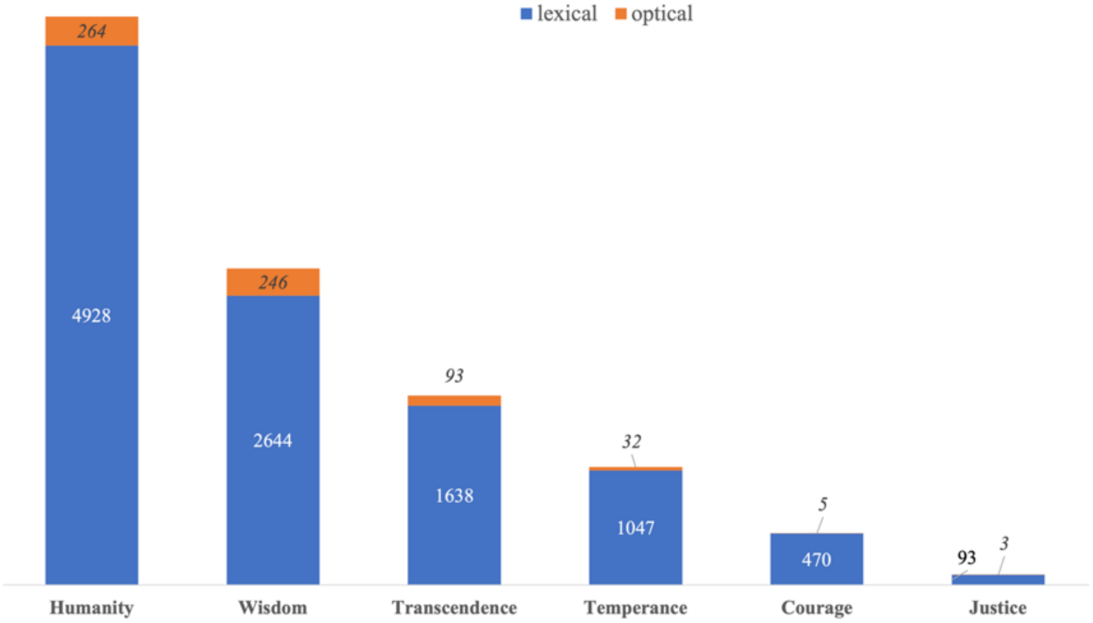
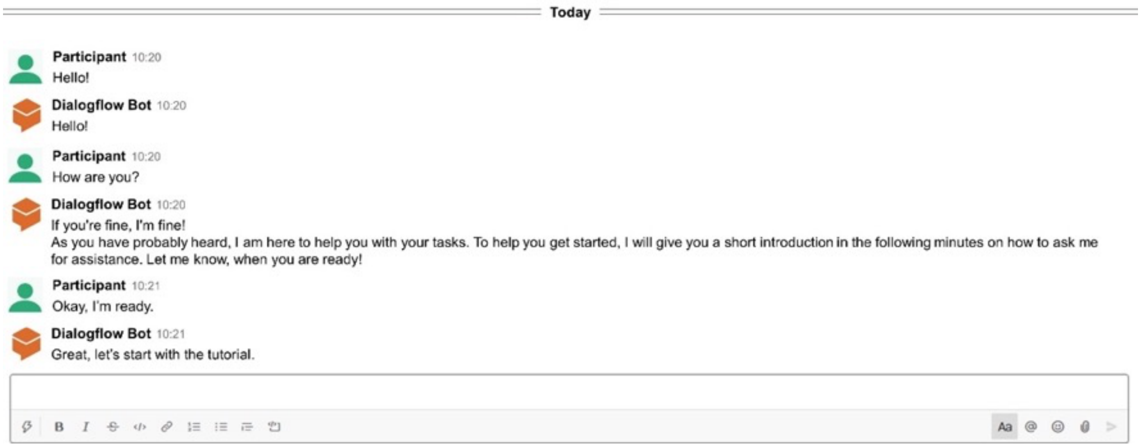


Figure 6. *Distribution of the virtues that were expressed by the 10 most successful virtual influencers on Instagram based on the findings of P8*

4.2 Effects on the Perception of Virtual Agents When They Take Over Human Tasks Related to Moral Responsibility

Building on social identity theory (Tajfel & Turner, 2004) and the concept of extended self (Belk, 2016), our results in P5 suggested that VAs can be perceived as tools and teammates at the same time (Mirbabaie, Stieglitz, Brünker, et al., 2021). With, on the one hand, an increase in the autonomy, automation, and capabilities of the systems and, on the other hand, almost realistic-looking avatars, for example, as virtual influencers, people attribute increasing moral responsibility to these systems. According to Mingers and Walsham (2010), it is important for an ethical discourse not only to provide guidelines for ethically correct actions but also to take into account the perceptions of those affected. Moreover, it is less attractive for companies to implement moral values and actions if the connection to their perception and particularly trust in responsible VAs is uncertain. Therefore, in this section, I provide an overview of identified concepts that are positively and negatively related to the perception of VAs and that explicitly or implicitly affect human-like trusting beliefs.

One factor that is important for trust between humans is a shared identity. Therefore, in P5, the author team examined the perception of a chatbot that provided assistance to a collaborative task in a virtual work context that the participants had to solve as part of a laboratory study. We assigned 50 participants to two groups: one group should solve the task assisted by the chatbot and one group was supported by a human chat partner. We drew conclusions on identification with the chatbot and the system as part of the participants' extended selves. On the one hand, our findings did not suggest that virtual collaboration with a chatbot, compared to a human chat partner, affects social identity. On the other hand, our findings indicated that people collaborating with chatbots identified less with their human teams after the interaction. Furthermore, we found a positive relationship between individuals' identification with the team and individuals' identification with technology. Figure 7 shows one exemplary conversation between a participant and the chatbot.



Participant 10:20
Hello!

Dialogflow Bot 10:20
Hello!

Participant 10:20
How are you?

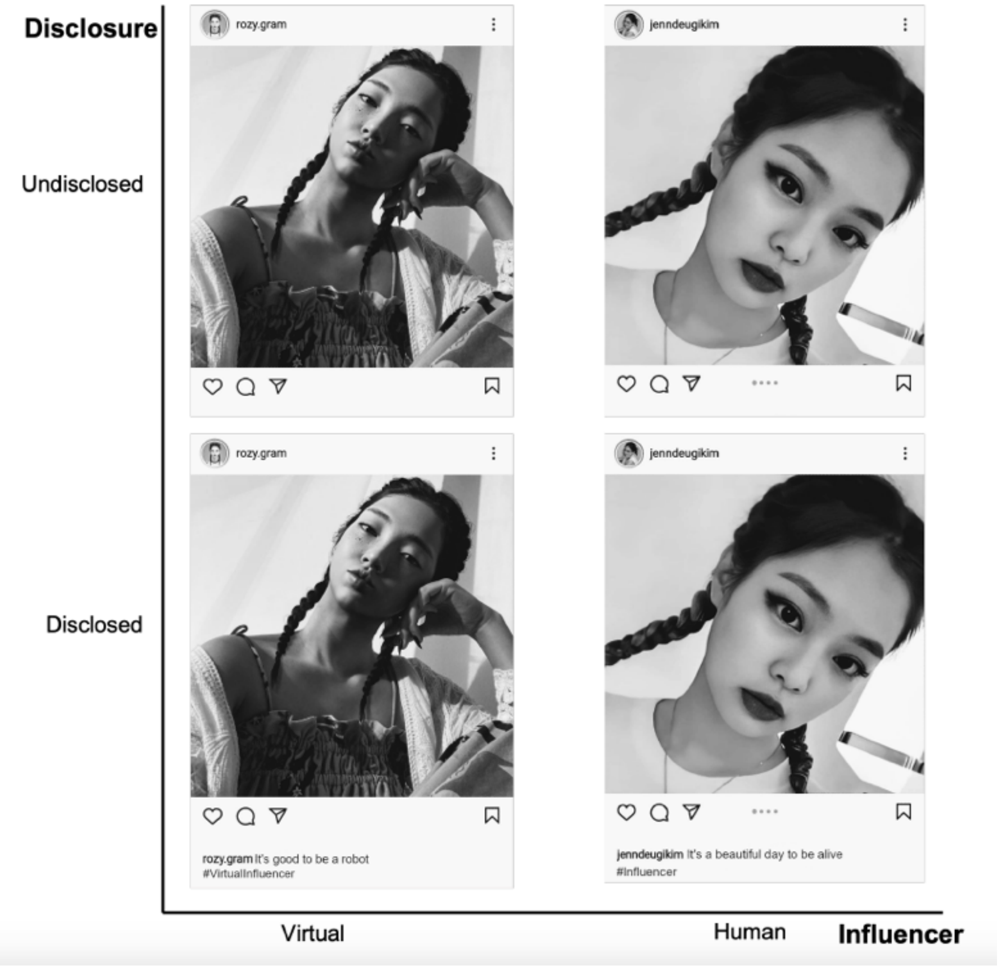
Dialogflow Bot 10:20
If you're fine, I'm fine!
As you have probably heard, I am here to help you with your tasks. To help you get started, I will give you a short introduction in the following minutes on how to ask me for assistance. Let me know, when you are ready!

Participant 10:21
Okay, I'm ready.

Dialogflow Bot 10:21
Great, let's start with the tutorial.

Figure 7. *Screenshot of one interaction between the chatbot and one participant from the study of P5*

Previous research suggests that people can perceive computers, especially VAs, as social actors (Lee & Nass, 2010; Nißen et al., 2022). Most of the empirical evidence is based on the self-reports of the study participants. In P6, the author team therefore not only based or study on self-reports but also directly measured the perception of highly human-like VAs by conducting a multimethod brain interface study. We combined survey data (N = 112) with neuronal data from a fNIRS laboratory study (N = 34) to be able to compare aspects such as human-like trust between human influencers and human-like virtual influencers on Instagram. In the third study of this paper, we also measured the perceived mind toward different human and virtual influencers and controlled for the effects of attractiveness, authenticity, familiarity, and ethnicity (N = 193). Our findings indicate that uncertainty in mind attribution toward virtual influencers was lower when the virtual influencers disclosed themselves as non-human. Although people often falsely classified virtual influencers as real humans, human-like trust toward the virtual influencers in our study was generally lower than trust in the presented human influencers. In addition, human influencers were rated higher in perceived social presence and lower in perceived uncanniness. Although the participants in Study 3 differed in their ethnicity, we did not find an effect of the ethnicity of the participants in comparison to the ethnicity of the influencers. We identified the attribution of the mind as a precedent mechanism for other aspects, such as trust, uncanniness, and social presence, which is not manipulated by perceived attractiveness and authenticity. Figure 8 shows an example of a human and a virtual influencer presented in this study.

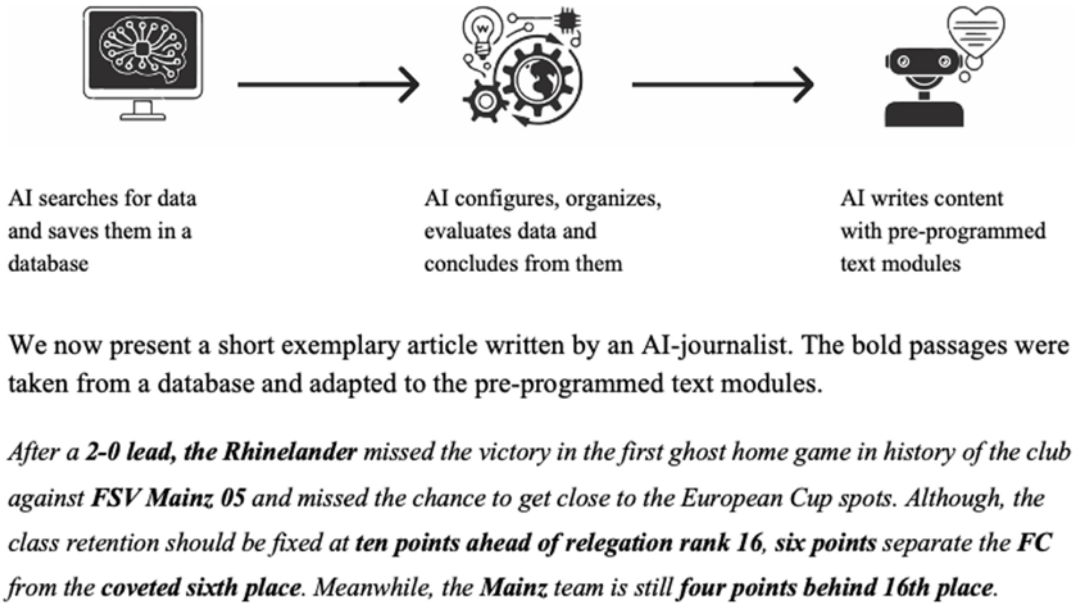


rozy.gram It's good to be a robot
#VirtualInfluencer

jenndeugikim It's a beautiful day to be alive
#Influencer

Figure 8. *Examples of the stimulus material presented in the study of P6*

One industry that is also highly affected by VAs is the media sector. In P7, we therefore examined the perception of content generated by an AI-based system in comparison to a human author. In an online survey study, we assigned 122 participants to four groups: two groups were presented with social media content that disclosed AI-authorship, whereas one of these groups received additional explanations of how the content was generated. The other two groups were presented with the same news content on different media organizations' websites and varied in their level of explanation. Although our findings did not suggest general differences between the perceived credibility of human- and AI-created content, AI-experienced users rated credibility lower than the unexperienced participants. The positive impact of the additional explanations provided in two of the groups showed an effect on the perceived credibility toward AI-generated content of organizations with a high-credibility disposition. Figure 9 shows the explanation provided for the AI-generated content.



AI searches for data and saves them in a database

AI configures, organizes, evaluates data and concludes from them

AI writes content with pre-programmed text modules

We now present a short exemplary article written by an AI-journalist. The bold passages were taken from a database and adapted to the pre-programmed text modules.

*After a **2-0 lead, the Rhinelander** missed the victory in the first ghost home game in history of the club against **FSV Mainz 05** and missed the chance to get close to the European Cup spots. Although, the class retention should be fixed at **ten points ahead of relegation rank 16**, **six points** separate the **FC** from the **coveted sixth place**. Meanwhile, the **Mainz** team is still **four points behind 16th place**.*

Figure 9. *Explanation for the participants of how the AI-based system generated the content for the study in P7*

Lastly, VAs are not limited to AI-based systems. One manifestation is virtual influencers that mimic human influencers on social media platforms, such as Instagram. In section 4.1, I have already summarized the findings of P8 on how they can express certain values and virtues. In P8, the author's team also considered the engagement rate and examined behavior, such as virtue signaling. We analyzed the content of the 10 most influential human-like virtual influencers and compared the expressed virtues with the values of the companies that used the virtual influencers to promote their products. We found a mismatch between the organizations' values and the expressed virtues of the virtual influencers that were not noticed by their followers. We think that virtual influencers trigger other aspects, such as uncertainty in mind attribution and uncanniness, mentioned in P6, which might distract from the focus on expressed values.

5 Discussion

Based on the results of the individual papers of this dissertation, insights can be derived on how organizations can allocate moral responsibility between humans and VAs to improve decision-making for individuals, groups, and society (RQ1). The interpretation of the results from P1, P2, P3, P4, and P8 is directed to experts, organizational decision-makers, and policymakers and provides explicit and implicit approaches for managing and allocating moral responsibility in different organizational contexts. They address the first research question and are discussed in section 5.1.

P1, P5, P6, P7, and P8 provide approaches for users interacting with and affected by VAs, which, according to Mingers (2010), is also essential when discussing moral principles and theories. The majority of these articles are specifically concerned with concepts affecting the perception of human-like trust toward VAs or the understanding of their general perception when VAs take over human tasks related to moral responsibility. These articles address RQ2 and are discussed in section 5.2.

5.1 How Organizations Allocate Moral Responsibility Between Human Users and Virtual Agents

The results from P1, P2, P3, and P4 contain different considerations for allocating moral responsibility between humans and VAs. The findings generally build on the idea of sharing moral responsibility between humans and machines (especially VAs), which was suggested by Johnson and Powers (2005), Verbeek (2011), and Behdadi and Munthe (2020). In contrast to these works, the results of the papers included in this dissertation, on the one hand, provide suggestions on how this allocation of moral responsibility can be established by organizations to achieve hybrid moral responsibility. Based on my findings, in this thesis, I define hybrid moral responsibility as a concept that involves the allocation of moral responsibility between humans and VAs when their collaborative decision-making exceeds the capabilities of either party acting alone, combining human ethical judgment with the computational abilities of VAs to make better moral choices that benefit potentially affected individuals, vulnerable groups, or society. In this allocation, humans take on the whole backward-looking moral responsibility, whereas both VAs and human users share the forward-looking responsibility. On the other hand, the findings present an overview of the effects on users' perceptions when VAs take over human tasks related to moral responsibility.

5.2 Virtual Agents Are Guides in Hybrid Moral Responsibility

Previous research suggested that in the development process, humans (mostly the developers themselves) are responsible for ensuring that VAs act according to the main bioethical principles of beneficence, non-maleficence, autonomy, justiciability, and explicability (Floridi et al., 2018) to achieve trustworthiness (Independent High-Level expert Group on Artificial Intelligence, 2019). However, during the interaction with the systems, the findings of the papers included in my dissertation suggest that forward-looking responsibility is increasingly transferred to the VA to ensure that users act morally and to guide them in their moral decision-making (see P1–P4). In P5, we found that although VAs supporting virtual collaboration in the workplace are mostly considered tools, they also simultaneously substitute for the role of a human collaborator. This extends the research on the role of machines as teammates (Seeber et al., 2020) by addressing the aspects of how VAs are perceived in hybrid work scenarios. One characteristic of a human collaborator is that of taking moral responsibility, which might indicate that people also attribute moral responsibility to VAs. This possible conclusion from P5 is also supported by Banks (2021) who stated that robots are associated with moral responsibility and that they more likely blamed for their behavior than human actors for their bad behavior. In contrast, the developers of ChatGPT struggle to prevent human users from requesting information for unethical actions (e.g., instructions on how to build a bomb) (Kington, 2022). In these cases, the manufacturer of the VA allocates the VA's task of preventing users from interactions that are morally questionable, even if they might continuously ask for such interactions. This, however, limits human users' autonomy, which suggests that there might be ethics principles that are more important than others in hybrid moral responsibility.

Although prior research concluded that people are more forgiving of moral mistakes made by other people than of moral mistakes made by VAs (Banks, 2021), people seem to follow their recommendations and outputs of VAs, even in ethical decision-making (see P1 and P4). The results in P4 indicate that VAs increasingly take over moral responsibility in healthcare by guiding physicians (e.g., in treatment decisions that will affect human lives). In the context of the ethical use of AI-based systems in hospitals, one interviewee from our study in P4 said clinicians would "rely on the technology and become dependent on it," and it would increasingly happen that "AI does the thinking and people act blindly" (E2). Our experts in P4 reported that one reason for this is time constraints and low technical knowledge. However, this seems to be contrary to research in the field of algorithm aversion, which assumes that domain experts are more likely to rely on human actors (Berger et al., 2021; Dietvorst et al., 2018; Kawaguchi, 2021). One explanation for this apparent contradiction could be that physicians cannot afford not to rely on VA recommendations due to a lack of time resources. Another explanation might be previous research suggesting that for complex moral decisions, ML tools, such as VAs, might be used to make a final decision in which there is only slight support from humans (Teodorescu et al., 2021). The results from P1, P2, P3, and P4 mainly emphasize the importance of humans as final decision makers with respect to moral decisions and allocated moral responsibility. Based on these findings, it can be concluded that organizations should allocate moral responsibility in such a way that VAs should function as advisers in moral decisions and take the lead in moral responsibility (see P1) but leave the final decision to these users (see P1, P2, P3, and P4).

Our findings suggest that VAs should question and filter the requests of their users for immoral actions (see P1 and P4). Furthermore, they should provide recommendations for moral actions and should actively discourage immoral behavior (see P1). However, humans are morally responsible for making the final decision (see P1 and P4) and are also proactively required to check the organization VA's compliance with the moral principles of beneficence, non-maleficence, justice, autonomy, and explicability by testing certain functions and their results (see P2 and P4).

The study in P1 also shows an approach of how the moral responsibility of a VA can not only provide recommendations toward more diversity in age, race, and gender of candidates on a hiring platform but can also provide explanations for these recommendations. However, our findings suggest that the role of explanations is highly dependent on context and content and is not the holy grail for pushing the effects of these recommendations (see P1). This underlines the complexity of the concept of XAI (Meske et al., 2022; Moradi & Samwald, 2021). A combination of local and global explanations might be a key component for achieving hybrid moral responsibility, but its effectiveness cannot be universally concluded. In contrast to knowledge from algorithm aversion and appreciation (Berger et al., 2021; Kawaguchi, 2021; Logg et al., 2019), domain knowledge of users does not seem to be an influencing factor for the effectiveness of explanations (see P1). For race discrimination, however, our findings for P1 showed an effect of the explanations. This is in line with Bigman et al. (2021), who found that emphasizing the threat of inequality because of race discrimination reduced aversion toward a VA. Our findings suggest that this might not be the case for all types of discrimination.

ChatGPT, the most successful VA in 2023 (Short & Short, 2023), and similar tools such as DALL-E 3[13] already experiment with limiting humans' autonomy to ensure non-

[13] https://openai.com/dall-e-3

maleficence and justice. For example, if a human user asks for content related to hate, threatening, self-harm, sex, minors, and violence, ChatGPT not only refuses the answer but also informs the user about the unethical nature of the request and provides certain advice.[14] Unfortunately, this allocation of moral responsibility to VAs does not correspond to any democratic or social participation process but is voluntary measures by the provider, which are mostly explained in an insufficient way. Other VAs, such as the FreedomGPT model, however, avoid assigning moral control mechanisms to their technologies by allowing users to request everything they want, including uncensored images and deepfakes (Newswire, 2023). Therefore, it is important that, on the one hand, democratic institutions provide detailed frameworks for allocating moral responsibility between humans and VAs, not the companies themselves. These guidelines should not be limited to regulations for VAs, such as the ethics guidelines for trustworthy AI provided by the European Union (European's Commission: HLEG, 2019) but should provide guidance on how organizations should allocate moral responsibility between users and VAs. Due to the findings from P2, on the other hand, it is important that human users interacting with VAs actively consent that their requests are beneficial for affected individuals, groups, or society and do not discriminate against vulnerable people or groups. Allocating moral responsibility, as described in this section, provides a general basis for achieving hybrid moral responsibility: moral decision-making that is better than that of a human or a VA when they are acting on their own.

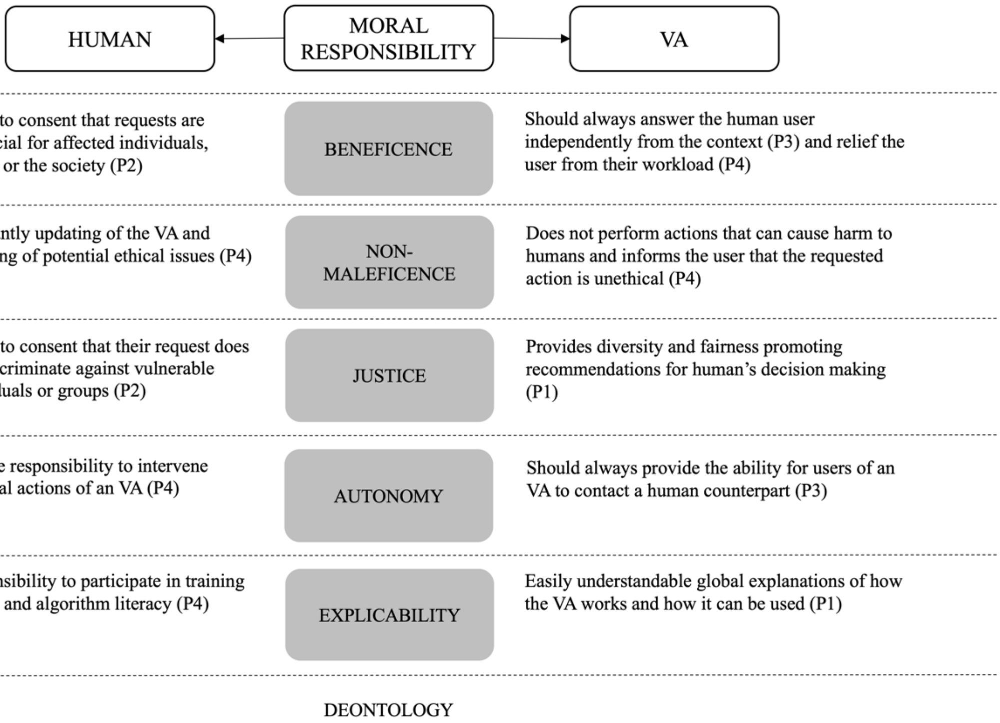
5.3 Organizations Can Learn from Different Normative Ethics Perspectives to Allocate Moral Responsibility

The articles included in this dissertation consider the allocation of moral responsibility to VAs and their users multidimensionally applying the three normative ethics perspectives of deontology, consequentialism, and virtue ethics (Chakrabarty & Erin Bass, 2013). These perspectives provide guidance for organizations to implement the allocation of moral responsibility. The results of this dissertation suggest that the deontological perspective is primarily focused on the definition of and compliance with rules for developers, providers of VAs, and its users. Clear ethical guidelines can be used to implement the deontological perspective (see P2). The consequentialism perspective is particularly relevant for assessing the impact of the use of VAs on individuals, groups, or society and the environment. This perspective is particularly important for ensuring trust in systems (see P9). The virtue ethics perspective can aim to equip human users, developers, and the system with virtues such as a sense of responsibility, compassion, and

[14] https://platform.openai.com/docs/guides/moderation/overview

ethical integrity. This could be implemented through training programs and ethical education (see P4).

Taking these three perspectives, Floridi's (2018) ethics principles are common ground for the consideration of how moral responsibility can be allocated, as they still provide a good framework for discussion in the current research discourse (Dwivedi et al., 2023). Therefore, organizations can implement the allocation of responsibility by considering any of Floridi et al.'s (2018) principle from a deontological, consequentialist, and virtue ethics perspective. Some aspects of this deontological perspective, such as the allocation of moral responsibility against the background of user autonomy and ensuring non-maleficence, have already been discussed in section 5.1. An overview of how the results of the papers included in this doctoral thesis can inform the exact allocation by principle from a deontological perspective, as shown in Figure 10. The content is derived from the findings, discussion, and implications sections of P1–4.



HUMAN	MORAL RESPONSIBILITY	VA
Needs to consent that requests are beneficial for affected individuals, groups or the society (P2)	BENEFICENCE	Should always answer the human user independently from the context (P3) and relief the user from their workload (P4)
Constantly updating of the VA and reporting of potential ethical issues (P4)	NON-MALEFICENCE	Does not perform actions that can cause harm to humans and informs the user that the requested action is unethical (P4)
Needs to consent that their request does not discriminate against vulnerable individuals or groups (P2)	JUSTICE	Provides diversity and fairness promoting recommendations for human's decision making (P1)
Has the responsibility to intervene immoral actions of an VA (P4)	AUTONOMY	Should always provide the ability for users of an VA to contact a human counterpart (P3)
Responsibility to participate in training of data and algorithm literacy (P4)	EXPLICABILITY	Easily understandable global explanations of how the VA works and how it can be used (P1)

Figure 10. *Summary of a deontological approach to allocating moral responsibility between humans and VAs*

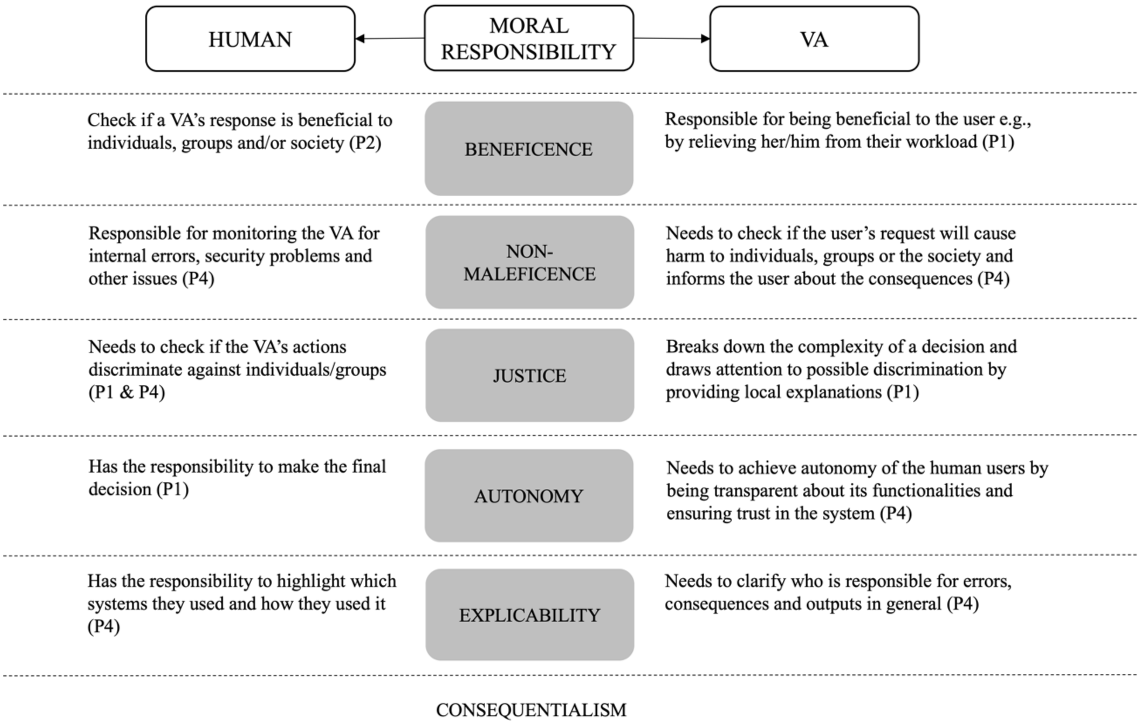
To answer RQ1, it is important not to limit the considerations of hybrid moral responsibility to one theoretical perspective of normative ethics. Some of the findings from P1, P2, P3, P4, and P8 can also be considered from a consequentialist perspective.

In P1, we provide an example of how VAs can help human resource (HR) managers achieve their goals (finding suitable candidates for a job) by providing concrete recommendations. According to Floridi et al. (2018), this can not only be seen as the fulfillment of the principle of justice but also includes the ethical principle of beneficence, as it benefits HR managers by providing support in difficult decisions. Furthermore, in P1, we provide one approach, which, on the one hand, showed that some participants acted immorally on a candidate selection platform and, for example, stated, “I would not invite a Turk” (ID 1405). This is an example of negative consequences and needs to be prevented by acting in accordance with the ethical principles of non-maleficence and justice. On the other hand, our results suggested that AI-based decision support toward more diversity in the hiring process increased the selection of older and female candidates as representatives for groups against which discrimination often occurs. Therefore, this can be seen as one possible approach to fulfilling these principles.

In general, the consequentialist aim of applying VAs, such as AI-based systems, in recruiting is not only to reduce the workload of HR managers but also to reduce discrimination in hiring (Ochmann & Laumer, 2019). Accordingly, some organizations do not fully entrust their employees to make the right moral decisions on their own.

If a VA is ascribed moral responsibility by users in the interaction, the question arises as to who is legally accountable for harm to individuals, groups, or society (Kordzadeh & Ghasemaghaei, 2021; Martin, 2019). Based on the results from P4, responsibility can be divided as follows: if third parties (in this case, patients) are affected by the interaction between VA and a user, the medical professional is morally responsible for informing them about the exact use and functioning of the systems to fulfill the principle of explicability and is liable for possible harm. If errors occur in the interaction and third parties are not directly affected, the system and thus the manufacturer can be held responsible. In P4, one interviewee said, “We as a company are accountable for keeping our stable clean. But we should also have the doctors who can question this again in case of doubt. But a certain amount of legal liability should also lie with the manufacturer, who should also be responsible for ensuring that the AI is always up to date” (E6). This could be transferred to other contexts, such as recruiting. A company using a VA in HR management is therefore morally responsible for making it transparent not only to the users but also to those affected by its outcome, which systems they used, and how they used them. The findings of P1 provide guidance on how the system can be explained to users. In contrast, the human using the system is responsible for which systems they use and how, as suggested by the findings in P4.

The overall allocation of moral responsibility derived from the papers in this thesis from a consequentialist perspective is shown in Figure 11.



HUMAN	MORAL RESPONSIBILITY	VA
Check if a VA's response is beneficial to individuals, groups and/or society (P2)	BENEFICENCE	Responsible for being beneficial to the user e.g., by relieving her/him from their workload (P1)
Responsible for monitoring the VA for internal errors, security problems and other issues (P4)	NON-MALEFICENCE	Needs to check if the user's request will cause harm to individuals, groups or the society and informs the user about the consequences (P4)
Needs to check if the VA's actions discriminate against individuals/groups (P1 & P4)	JUSTICE	Breaks down the complexity of a decision and draws attention to possible discrimination by providing local explanations (P1)
Has the responsibility to make the final decision (P1)	AUTONOMY	Needs to achieve autonomy of the human users by being transparent about its functionalities and ensuring trust in the system (P4)
Has the responsibility to highlight which systems they used and how they used it (P4)	EXPLICABILITY	Needs to clarify who is responsible for errors, consequences and outputs in general (P4)

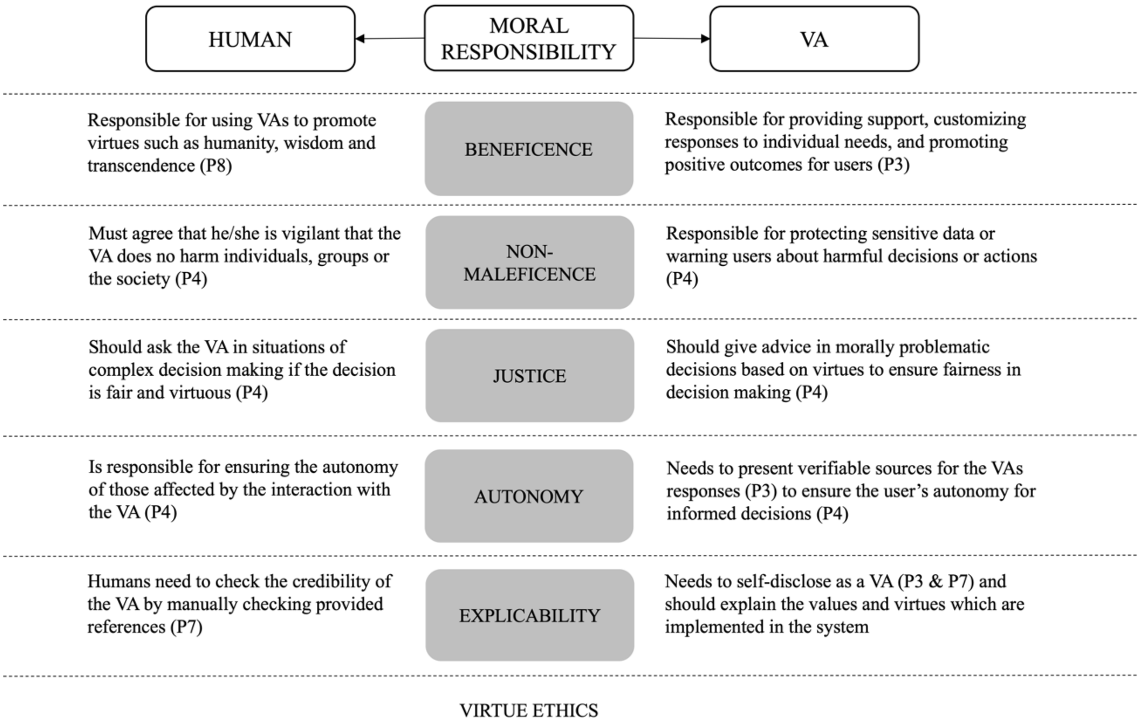
Figure 11. *Summary of a consequentialist approach to allocating moral responsibility between humans and VAs*

Due to the black box character of AI-enabled VAs (Berente et al., 2021; Guidotti et al., 2018), it is almost impossible to achieve a clear allocation of moral responsibility based solely on a deontological set of rules or a focus on consequences, which are often unpredictable in the interactions between humans and VAs. Therefore, the findings of P4 and P8 also aim to inform a virtue ethics consideration to achieve hybrid moral responsibility.

As not all VAs are AI-based and fully automated, P8's findings suggest that humans using them should focus on promoting virtues such as humanity, wisdom, and transcendence. According to P8, these were found to be the virtues most frequently expressed by the most influential virtual influencers. To implement such virtues into more automated systems, they can be built on the suggestions and approaches of Stenseke (2021), who implemented the virtues of courage, generosity, and honesty. In practice, these virtues can be expressed by VAs in customized responses, as suggested in P3, in the context of crisis communication.

According to the findings of P3 and P7, the virtuousness of VAs also requires systems to self-disclose as non-human actors and explains the values implemented in the system. This constitutes an extension of explanability research (e.g., Meske et al., 2022; Moradi & Samwald, 2021) to include the facet of explaining ethical values. The systems' moral responsibility to provide explanations is also closely linked to the principle of autonomy. The findings of P3 and P4 suggest that a VA should provide verifiable sources for responses to the users to ensure informed decision-making.

Human users, in contrast, are responsible for checking the credibility of a VA by checking provided references (P7) and are vigilant that the system does not harm individuals, groups, or society (P4). An overview of a virtue ethics perspective for achieving hybrid moral responsibility is shown in Figure 12.



HUMAN	MORAL RESPONSIBILITY	VA
Responsible for using VAs to promote virtues such as humanity, wisdom and transcendence (P8)	BENEFICENCE	Responsible for providing support, customizing responses to individual needs, and promoting positive outcomes for users (P3)
Must agree that he/she is vigilant that the VA does no harm individuals, groups or the society (P4)	NON-MALEFICENCE	Responsible for protecting sensitive data or warning users about harmful decisions or actions (P4)
Should ask the VA in situations of complex decision making if the decision is fair and virtuous (P4)	JUSTICE	Should give advice in morally problematic decisions based on virtues to ensure fairness in decision making (P4)
Is responsible for ensuring the autonomy of those affected by the interaction with the VA (P4)	AUTONOMY	Needs to present verifiable sources for the VAs responses (P3) to ensure the user's autonomy for informed decisions (P4)
Humans need to check the credibility of the VA by manually checking provided references (P7)	EXPLICABILITY	Needs to self-disclose as a VA (P3 & P7) and should explain the values and virtues which are implemented in the system

Figure 12. *Summary of a virtue ethics approach to allocating moral responsibility between humans and VAs*

Implementing the practices shown in Figure 12 and the previous figures can help organizations make better moral decisions than their employees or their VAs could make individually. Which perspectives are appropriate for an organization depends on the industry, corporate philosophy, and objectives. In general, the deontological perspective is more appropriate for companies that prefer clear guidelines and rules (for example, because they operate in highly regulated industries, such as healthcare). The consequentialism approach is more suitable for companies with a focus on social responsibility (e.g., emergency response agencies). Virtue ethics is suitable for organizations with a strong corporate culture and ethical leadership as well as a lot of individual freedom and personal responsibility (e.g., universities or consulting firms).

5.4 Factors Influencing the Perception of Virtual Agents That Take Over Human Tasks Related to Moral Responsibility

To answer RQ2, in P1, P5, P6, P7, and P8, my coauthors and I identified and discussed important concepts affecting the perception of VAs that take over human tasks related to moral responsibility. As the first concept that is relevant in the perception of VAs, the findings of P1 suggested that explicability, including an explanation of the system and its outputs, is a highly important concept whose effectiveness highly depends on the content and context. Extending the XAI quality criteria of Meske et al. (2022), the findings of P1 and P7 suggest that content quality is an additional factor that impacts the perception of VAs.

As VAs are increasingly capable of complementing humans in virtual collaboration, the findings from P5 identified the concept of virtually extended identification as another relevant concept in the perception of VAs. This supports the assumption that a shared identification is no longer limited to trust between humans, as originally stated by Lewicki and Bunker (1997, 1996) but is also relevant to human-like trust toward VAs. Although trust was not directly at the center of P5, the dual role of the VAs being perceived as supportive tools and social actors parallels human-like and system-like trust categories (Lankton et al., 2015).

VAs not only increase their level of sophistication. The findings of P6 suggest that many people falsely classify virtual influencers as real humans. This informs the debate on whether computer-generated avatars are about to cross the uncanny valley (Seymour et al., 2021) by showing that uncanniness is still rated higher for virtual influencers and social presence is rated higher for similar-looking human influencers. However, although people often falsely classified virtual influencers as real humans, virtual influencers were perceived as lower in human-like trust. Moreover, we found indications of a conflict in mind attribution when perceiving virtual influencers based on self-reports and neuronal data. This builds on previous research suggesting that people try to attribute a mind to artificial agents (Vinanzi et al., 2019), which also provides evidence for a two-system account, assuming that mind attribution is based on an implicit and explicit process (Meinhardt-Injac et al., 2018). For perceiving virtual influencers, our findings suggest a conflict in the implicit system due to uncertainty. Our findings in P6 further indicate that this conflict in mind attribution might be solved by disclosing whether an influencer is human or purely virtual.

Disclosure is, therefore, another important concept that can affect the perceptions of VAs. The findings of P6 suggest that disclosure impacts perceived uncanniness and social presence but does not affect trust in the systems. Based on the findings of P7, the credibility of the organization that provides the VA impacts trust independently of the abilities of the system. This indicates that external factors, such as an organization's credibility, weigh more than system-like trusting beliefs, such as reliability, helpfulness, and functionality (Mcknight et al., 2011). Explanations further strengthen the perception of trust in such systems of organizations with high credibility.

A further concept that could be identified by P7 was virtual experience amplification. Our findings suggest that people with a higher level of experience with virtual technologies, such as social media or AI-based systems, rated the text generated by VAs as lower in trust than people with less experience. This supports similar findings on algorithmic aversion (Bigman et al., 2021; Dietvorst et al., 2015) and algorithmic appreciation (Logg et al., 2019; You et al., 2022), suggesting that domain experts are more likely to choose a human forecaster and laypeople are more likely to choose an algorithm when they have the choice between those two.

Lastly, P8's findings suggest that organizations use cooperation with virtual influencers to promote certain values and virtues. This can be misused for virtue signaling when virtual influencers express virtues that stand in contrast to the moral actions of a company for which they are promoting. Our findings revealed that this practice was not noticed by the influencers' followers and did not result in negative consequences for the companies, such as a decrease in trust or credibility. Previous research has already warned about the negative consequences of using virtual influencers (Robinson, 2020). Therefore, it is important to derive a better understanding of the perception of VAs. In Table 7, I provide an overview of the concepts identified, analyzed, and discussed in P1, P5, P6, P7, and P8 and describe the effects on the perception of VAs.

Table 7. *Concepts that influence the individual perception of and especially human-like trust in VAs*

Concept	Effect on the perception of VAs	Paper
Explicability in the context of VAs	Although explicability in the context of VAs is a highly important principle from an ethical point of view, it often does not affect trust in the system. One reason for this might be that trust in VAs is already very similar to trust in other humans. However, in some contexts (e.g., on social media) and for some types of explanations, our findings suggest slight differences in trust perception.	P1 and P7
Virtually extended identification with VAs	The process of extending one's self by comparing one's self with the VA. It further describes the matching of one's identity with the perceived identity of the VA. At the same time, the capabilities of the system can be seen as an extension of one's own capabilities. The agents thus pursue a dual role that is not mutually exclusive. Virtually extended identification is a key component for understanding collaboration with VAs.	P5

Conflict in mind attribution toward VAs	Regardless of whether someone or something has a mind, people try to attribute a mind to many things. Mind attribution toward VAs can be either a slower explicit cognitive-reflective process or a faster implicit social-perceptual process. When people are unsure whether a counterpart is human or not (e.g., when a VA looks or acts human-like), there can be a conflict in mind attribution, which leads to uncertainty, lower trust, less social presence, and more uncanniness.	P6
(Self-)disclosure of VAs	Self-disclosure does not affect human-like trust and perceived social presence toward VAs but results in a higher level of perceived uncanniness. This can be explained by the fact that disclosing VAs as artificial agents might mitigate a decision conflict in mind attribution, which facilitates uncanniness judgment.	P6
The interplay of source credibility and explanations	Independently from the real abilities of VAs, people rate those systems higher in trust that are provided by an organization which they perceive as credible. Furthermore, for organizations perceived as credible, explanations for their VAs further increase trust in the information provided by the systems.	P7
Virtual experience amplification	Experience with technologies such as social media or AI-based systems negatively influences trust toward VAs. This may be related to an increasing awareness of the potential negative consequences of working with these technologies.	P7
Virtue signaling of VAs	Virtue signaling, which is the practice of expressing certain moral values to a targeted audience with the aim of convincing others of one's moral integrity, can be used by organizations through VAs. On the one hand, it can be used to communicate certain organizational values in a natural way, as expressing virtues generates a high amount of user engagement. On the other hand, it can be misused when VAs communicate values that are contrary to the behavior of the organization.	P8

These concepts and effects provide invaluable insights into the complex dynamics that influence the perception of VAs and the formation of trust. Understanding the nuances of explainability, virtual identification, thought attribution conflict, self-disclosure, source credibility, virtual experience amplification, and virtue signaling not only enriches the understanding of human–VA interaction in IS research but also provides a strategic compass for organizations aiming to design, deploy, and maintain VAs that promote trust, minimize uncertainty, align with organizational values, and effectively engage users within an ethical framework. These concepts can guide organizations that decide to implement hybrid moral responsibility by considering one or more normative ethics perspectives to ensure employees' and customers' well-being and satisfaction when interacting with such a VA.

6 Conclusion

To examine how organizations can allocate moral responsibility between human users and VAs to achieve hybrid moral responsibility, I conducted different studies that took three theoretical lenses from normative ethics. I applied each lens to the principles of beneficence, non-maleficence, autonomy, justice, and explicability. Furthermore, I examined the effects on the perception of VAs when they take over human tasks related to moral responsibility. I identified concepts and factors that impact the perception, especially human-like trusting beliefs. Both the derived knowledge on allocated hybrid moral responsibility and the identified concepts impacting the perception of organizational virtual influencers provide implications for research and practice.

6.1 Contribution to Research

The results of this dissertation contribute to IS research and related disciplines, such as psychology, computer science, and social science, observing the perspectives of experts, decision-makers, and policymakers who manage the use of VAs and of those who interact with these systems or who might be affected by their use. I therefore revealed how a discourse ethical approach (Mingers & Walsham, 2010) can be followed for addressing a problem that is of high relevance for ethics research in the IS discipline.

Answering RQ1, I, on the one hand, provide an understanding for experts, decision-makers, and policymakers on which moral responsibilities can be taken over by VAs to provide relief to human users and minimize ethical issues in their interactions with human employees or customers. VAs can function as advisors for human users to raise their awareness of ethical behavior. Raising the awareness of the moral responsibility of VAs is important, as humans tend to underestimate how much their ethical decision-making is influenced by VAs (Krügel et al., 2023). On the other hand, I provide insights into which responsibilities need to be allocated to human users to update and control VAs. This mutual control can result in a higher level of moral behavior than both a VA and human users could have achieved on their own, which I call hybrid moral responsibility. This builds on and contributes to the concept of hybrid intelligence (Dellermann et al., 2019) by providing insights into how the complex problem of moral issues in the interaction between humans and VAs and in the impact of their interaction can be addressed collectively with superior results.

Furthermore, this work contributes to IS research by providing an approach to applying the three normative ethics perspectives of deontology, consequentialism, and virtue ethics to the ethics principles of beneficence, non-maleficence, autonomy, justice, and explicability in the context of the interactions between humans and VAs. I thereby contribute by addressing the problem and closing the gap in philosophical debates on whether moral responsibility can be delegated to non-human actors (Behdadi & Munthe, 2020) and IS research on the role of managing VAs such as AI-based systems in human–computer interaction and augmentation (Seeber et al., 2020; Teodorescu et al., 2021).

Answering RQ2, this work contributes to a better understanding of the perception of VAs by applying knowledge from psychological theories such as ToM (Frith & Frith, 2010; Mitchell, 1997) and social identity theory (Brown, 2000; Tajfel & Turner, 2004) to IS contexts and extending these theories by discussing phenomena such as virtually extended identification or a conflict in mind attribution. By answering RQ2, I further contribute to research by transferring concepts discussed in the IS discipline such as XAI (Förster et al., 2020; Meske et al., 2022) and resistance to IT change (Krovi, 1993; Laumer & Eckard, 2010) to the perception of humans' interactions with VAs to understand their positive and negative impacts, especially perceived trust toward these systems. Moreover, this work contributes to IS research by considering further concepts, such as virtue signaling (Wallace et al., 2020), to, on the one hand, provide an understanding of how VAs can be used by organizations in accordance with their values and, on the other hand, how expressing certain virtues and values is perceived by their customers.

6.2 Contribution to Practice

This dissertation contributes to practice by considering different contexts, such as healthcare, journalism, or crisis communication, and by providing design guidelines for developing and applying VAs for interaction with customers and employees. I provide concrete guidance on how managers need to address the concerns of their employees when VAs, such as AI-based systems, are introduced in the workplace. Furthermore, I derived principles for the design of such systems. The knowledge derived in this dissertation can also be used by organizations to train their employees in using VAs in accordance with ethical norms and policies and to improve their data and algorithm literacy. This work also reveals why it is important for organizations and their management to explain the use of VAs not only to their own employees but also to their customers.

Furthermore, the findings of this dissertation can be used to design work environments in which VAs and humans work together to achieve superior results collectively. This can, on the one hand, relieve their employees from repetitive work and augment them in complex decision-making, which might have positive effects on satisfaction and work efficiency. On the other hand, the results can support organizations by avoiding ethical problems toward individuals, groups, or society, which could have negative effects on their reputations.

In some contexts, such as healthcare, we provide concrete guidelines on how interactions between humans, such as physicians or patients, and VAs can be designed. In another context, such as hiring, we contribute by highlighting how HR managers and VAs can collaborate to ensure the recruitment of the best suitable candidate and to avoid discrimination due to demographic attributes, such as gender, race, or age. In the context of crisis communication, the findings of this dissertation reveal how important it is to use VAs to answer crisis-related questions and forward requests that cannot be processed by the system to a human encounter. Furthermore, we provide insights into how expressing virtues and values by applying VAs, such as virtual influencers, can result in a high user engagement rate.

6.3 Limitations

This work has some limitations. Allocating moral responsibility between humans and VAs does not eliminate immoral behavior and consequences that cause harm to individuals, groups, or society. It can contribute to minimizing ethical issues, but both humans and VAs can still behave immorally due to unidentified biases or by the lack of human will.

The proposed allocation of moral responsibility is based on eight empirical studies. Therefore, the generalizability of the results is limited. I focused on specific application domains, such as the hiring process, crisis communication, healthcare, journalism, online marketing, and virtual teamwork. However, no conclusion can be drawn regarding the allocation of moral responsibility in other contexts, such as education or research. In this work, the focus lies on the IS discipline as the primary audience, embracing an interdisciplinary viewpoint encompassing VAs and user perceptions. The author meticulously examined findings from various other fields, such as psychology, cognitive science, and communication science. However, due to the primary objective of achieving hybrid moral responsibility and examining its perception, this dissertation and its related papers could not delve extensively into the concepts and theories utilized (e.g., XAI, ToM, or social and IT identity).

Moreover, most of the studies were conducted in an experimental environment to control confounding variables. This allowed us to investigate influencing factors on the perception of VAs but might be contrary to their perceptions in more natural environments. Cultural background can also play an important role in the perception of VAs. Although we controlled for the influence of different ethnicities in P6, this was not the case in all the studies in this dissertation. Furthermore, the perception of VAs, especially the formation of trust, is a long process. However, we did not conduct any longitudinal studies in this dissertation. Thus, it cannot be concluded how the perception of VAs performing human tasks related to moral responsibility changes over time.

Lastly, the underlying technologies involved in VAs are constantly and rapidly changing. One example is that generative AI technologies, such as ChatGPT, or resulting AI companions and copilots, had not even been released at the time the studies in this thesis were conducted. Although we continuously updated the literature in this synopsis, the literature reviews conducted as part of this dissertation do not reflect the latest developments in research and practice.

6.4 Future Research

Previous scholars need to build on the conclusions of this work and extend the knowledge of how moral responsibility can be allocated between humans and VAs by considering recent technological advancements and deepening the effect of concepts such as XAI on the perception of such systems. One starting point could be the examination of the effectiveness of other more technical XAI approaches than that from P1 in order to derive knowledge on more diverse decision-making to achieve hybrid moral responsibility. Although previous research has suggested sharing moral responsibility between humans and VAs, there is still ongoing debate on whether non-human actors can take moral responsibility. Future IS scholars need to check whether there are new conclusions to this debate in disciplines such as philosophy.

Furthermore, moral responsibility can involve different types. Therefore, future research needs to be aware of concepts such as forward- and backward-looking responsibilities. One approach to implementing this could involve a more detailed examination of the perception of a VA taking forward-looking responsibility in comparison to one taking backward-looking or role-based responsibility. This might contribute to our understanding of the role a VA can play in the future in an organization and society.

Future research also needs to consider the concept of hybrid moral responsibility and test how the implementation of suggestions of this work regarding the allocation will improve moral decision-making and benefit individuals, groups, and society more than humans and VAs do individually. This could be tested in controlled environments, such as laboratories or online experiments. Future research should also examine suggestions for the allocation of hybrid moral responsibility in more real-world scenarios by conducting case studies and designing science and action research to ensure its applicability to practice. Scholars can build on the guidelines (e.g., from P2 and P4) and design principles (P3) provided in this work.

By examining more contexts, such as education and manufacturing, the generalizability of the findings of this work can be strengthened. For this, technologies such as ChatGPT in education or responsibilities in autonomous driving would provide valuable extensions to the understanding of hybrid moral responsibility. It would also be highly valuable to examine the perception of tools, such as Microsoft 365 or the GitHub copilot, which are particularly designed to provide guidance for human employees as a simulated team colleague. Valuable insights would involve checking whether people attribute moral responsibility to these VAs and how they perceive moral advice from these agents. Another promising approach would be the consideration of the effect of virtually extended identification on human-like trust and testing systems, which allow allocating moral responsibility toward their perception and acceptance.

Although debates on regulating generative AI technologies, such as ChatGPT, are important for preventing harm to individuals, vulnerable groups, or society, they often neglect the fact that they are used in hybrid scenarios, and their outputs are produced by an interplay of both human and VA inputs. The findings of this dissertation provide initial suggestions on how mutual support and control between those actors (which is not limited to human oversight but also includes system recommendations for moral actions) contribute to maximizing positive outcomes for individuals, vulnerable groups, and society that both humans and VAs could not achieve individually.

References

Aleksander, I. (2017). Partners of humans: A realistic assessment of the role of robots in the foreseeable future. *Journal of Information Technology*, *32*(1), 1–9. https://doi.org/10.1057/s41265-016-0032-4

Alexander, L., & Moore, M. (2007). *Deontological ethics*. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/Ethics-deontological/

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence*, *12*(3), 251–261. https://doi.org/10.1080/09528130050111428

Alter, S. (2022). Understanding artificial intelligence in the context of usage: Contributions and smartness of algorithmic capabilities in work systems. *International Journal of Information Management*, *67*(August). https://doi.org/10.1016/j.ijinfomgt.2021.102392

Applebaum, S. H. (1997). Socio-technical systems theory: an intervention strategy for organizational development. *Management Decision*, *35*(6), 452–463. https://doi.org/10.1108/00251749710173823

Arsenyan, J., & Mirowska, A. (2021). Almost human? A comparative case study on the social media presence of virtual influencers. *International Journal of Human Computer Studies*, *155*(June), 102694. https://doi.org/10.1016/j.ijhcs.2021.102694

Aversa, P., Cabantous, L., & Haefliger, S. (2018). When decision support systems fail: Insights for strategic information systems from Formula 1. *Journal of Strategic Information Systems*, *27*(3), 221–236. https://doi.org/10.1016/j.jsis.2018.03.002

Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of Moral Education*, *31*(2), 101–119. https://doi.org/10.1080/0305724022014322

Banks, J. (2021). Good Robots, Bad Robots: Morally Valenced Behavior Effects on Perceived Mind, Morality, and Trust. *International Journal of Social Robotics*, *13*(8), 2021–2038. https://doi.org/10.1007/s12369-020-00692-3

Barocas, S., & Selbst, A. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3). https://doi.org/10.15779/Z38BG31

Batista, A., & Chimenti, P. (2021). " Humanized Robots ": A Proposition of Categories to Understand Virtual Influencers. *Australasian Journal of Information Systems*, *25*, 1–27.

Bawa, A., Khadpe, P., Joshi, P., Bali, K., & Choudhury, M. (2020). Do Multilingual Users Prefer Chat-bots that Code-mix? Let's Nudge and Find Out! *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1–23. https://doi.org/10.1145/3392846

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of Biomedical Ethics* (8th ed.). Oxford University Press.

Behdadi, D., & Munthe, C. (2020). A Normative Approach to Artificial Moral Agency. *Minds and Machines*, *30*(2), 195–218. https://doi.org/10.1007/s11023-020-09525-8

Belk, R. (2016). ScienceDirect Extended self and the digital world. *Current Opinion in Psychology*, *10*, 50–54. https://doi.org/10.1016/j.copsyc.2015.11.003

Benbasat, I., & Wang, W. (2005). Trust In and Adoption of Online Recommendation Agents. *Journal of the Association for Information Systems*, *6*(3), 72–101. https://doi.org/10.17705/1jais.00065

Benlian, A., Wiener, M., Cram, W. A., Krasnova, H., Maedche, A., Möhlmann, M., Recker, J., & Remus, U. (2022). Algorithmic Management: Bright and Dark Sides, Practical Implications, and Research Opportunities. *Business and Information*

Systems Engineering, *64*(6), 825–839. https://doi.org/10.1007/s12599-022-00764-w

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. *Business and Information Systems Engineering*, *63*(1), 55–68. https://doi.org/10.1007/s12599-020-00678-5

Bigman, Y. E., Yam, K. C., Marciano, D., Reynolds, S. J., & Gray, K. (2021). Threat of racial and economic inequality increases preference for algorithm decision-making. *Computers in Human Behavior*, *122*(April). https://doi.org/10.1016/j.chb.2021.106859

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *30th Conference on Neural Information Processing Systems (NIPS2016)*.

Brachten, F., Kissmer, T., & Stieglitz, S. (2021). The acceptance of chatbots in an enterprise context – A survey study. *International Journal of Information Management*, *60*(June), 102375. https://doi.org/10.1016/j.ijinfomgt.2021.102375

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. *International Conference on Internet Science*, 377–392. https://doi.org/10.1007/978-3-319-70284-1_30

Brown, R. (2000). Social identity theory: Past achievements, current problems and future challenges. *European Journal of Social Psychology*, *30*(6), 745–778. https://doi.org/10.1002/1099-0992(200011/12)30:6<745::AID-EJSP24>3.0.CO;2-O

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at Work. *NATIONAL BUREAU OF ECONOMIC RESEARCH*, *Working Pa*(31161). http://www.nber.org/papers/w31161

Bussler, F. (2020). *Will The Latest AI Kill Coding?* Towards Data Science. https://towardsdatascience.com/will-gpt-3-kill-coding-630e4518c04d

Canhoto, A. I., & Clear, F. (2020). Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential. *Business Horizons*, *63*(2). https://doi.org/10.1016/j.bushor.2019.11.003

Carter, M., & Grover, V. (2015). Me, my self, and I(T): Conceptualizing information technology identity and its implications. *MIS Quarterly: Management Information Systems*, *39*(4), 931–958. https://doi.org/10.25300/misq/2015/39.4.9

Cecez-Kecmanovic, D. (2011). Doing critical information systems research-arguments for a critical research methodology. *European Journal of Information Systems*, *20*(4), 440–455. https://doi.org/10.1057/ejis.2010.67

Chakrabarty, S., & Erin Bass, A. (2013). Comparing Virtue, Consequentialist, and Deontological Ethics-Based Corporate Social Responsibility: Mitigating Microfinance Risk in Institutional Voids. *Journal of Business Ethics*, *126*(3), 487–512. https://doi.org/10.1007/s10551-013-1963-0

Chedrawi, C., & Howayeck, P. (2019). Artificial Intelligence a Disruptive Innovation in Higher Education Accreditation Programs: Expert Systems and AACSB. *Lecture Notes in Information Systems and Organisation*, *30*, 115–129. https://doi.org/10.1007/978-3-030-10737-6_8

Chomanski, B. (2023). A Moral Bind? — Autonomous Weapons, Moral Responsibility, and Institutional Reality. *Philosophy and Technology*, *36*(2), 1–14. https://doi.org/10.1007/s13347-023-00647-2

Cotter, K. (2021). “Shadowbanning is not a thing”: black box gaslighting and the power to independently know and credibly critique algorithms. *Information Communication and Society*, *26*(6), 1226–1243. https://doi.org/10.1080/1369118X.2021.1994624

Dastin, J. (2022). *Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women*. Ethics of Data and Analytics. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business and Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Diederich, S., Brendel, A.B., Lichtenberg S., and Kolbe, L. M. (2019). Design for fast request fulfillment or natural interaction? Insights from an experiment with a conversational agent. *European Conference on Information Systems*, 1–17.

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2019). On Conversational Agents in Information Systems Research: Analyzing the Past to Guide Future Work. *Proceedings of the International Conference on Wirtschaftsinformatik*.

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2020). Designing Anthropomorphic Enterprise Conversational Agents. *Business and Information Systems Engineering*, *62*(3), 193–209. https://doi.org/10.1007/s12599-020-00639-y

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dignum, V. (2019). Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. In *Arizona State Law Journal* (1st ed., Vol. 51). Springer Cham. https://doi.org/https://doi.org/10.1007/978-3-030-30371-6

Doherty, N. F., & King, M. (2005). From technical to socio-technical change: Tackling the human and organizational aspects of systems development projects. *European Journal of Information Systems*, *14*(1), 1–5. https://doi.org/10.1057/palgrave.ejis.3000517

Doshi, R. H., Bajaj, S. S., & Krumholz, H. M. (2023). ChatGPT: Temptations of Progress. *The American Journal of Bioethics*, *23*(4), 6–8. https://doi.org/10.1080/15265161.2023.2180110

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management*, *48*(January), 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Dunn, J. (2000). Trust and Political Agency. In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (electronic, pp. 73–93). Department of Sociology, University of Oxford. https://doi.org/10.1515/9781400861453.26

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*(March). https://doi.org/10.1016/j.ijinfomgt.2023.102642

Ebel, P., Söllner, M., Leimeister, J. M., Crowston, K., & de Vreede, G.-J. (2021). Hybrid intelligence in business networks. *Electronic Markets*, *31*(2), 313–318. https://doi.org/10.1007/s12525-021-00481-4

Elkin-Koren, N. (2020). Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. *Big Data and Society*, *7*(2). https://doi.org/10.1177/2053951720932296

Esmaeilzadeh, P. (2021). How does IT identity affect individuals' use behaviors associated with personal health devices (PHDs)? An empirical study. *Information and Management*, *58*(1), 103313. https://doi.org/10.1016/j.im.2020.103313

European's Commission: High Level Expert Group (HLEG). (2019). *Ethics guidelines for trustworthy AI Shaping Europe's digital future*. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Fahlquist, J. N. (2009). Moral responsibility for environmental problems - Individual or institutional? *Journal of Agricultural and Environmental Ethics*, *22*(2), 109–124. https://doi.org/10.1007/s10806-008-9134-5

Farina, L. (2022). Sven Nyholm, Humans and Robots; Ethics, Agency and Anthropomorphism. *Journal of Moral Philosophy*, *19*(2), 221–224. https://doi.org/https://doi.org/10.1163/17455243-19020007

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, *132*, 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009 Please

Fieser, J. (2018). *Internet Encyclopedia of Philosophy*. Ethics. https://web.archive.org/web/20180119084940/http://www.iep.utm.edu/ethics

Floridi, L. (2016). Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, *1*(6), 261–262. https://doi.org/10.1038/s42256-019-0055-y

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020). Fostering human agency: A process for the design of user-centric xai systems. *International Conference on Information Systems, ICIS 2020*, 0–17.

Friedman, B., Kahn, P. H., & Howe, D. C. (2000). Friedmann, Kahn, Howe (2000) trust online. *Communication of the ACM*, *43*(12), 34–40.

Frith, U., & Frith, C. (2010). The social brain: Allowing humans to boldly go where no other species has been. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1537), 165–175. https://doi.org/10.1098/rstb.2009.0160

Ghandeharioun, A., McDuff, D., Czerwinski, M., & Rowan, K. (2019). EMMA: An Emotion-Aware Wellbeing Chatbot. *8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019*, 15–21. https://doi.org/10.1109/ACII.2019.8925455

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goldkuhl, G. (2012). Pragmatism vs interpretivism in qualitative information systems research. *European Journal of Information Systems*, *21*(2), 135–146.

https://doi.org/10.1057/ejis.2011.54

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5). https://doi.org/10.1145/3236009

Gulati, S. N., Sousa, S. C., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour and Information Technology*, *38*(10), 1004–1015. https://doi.org/10.1080/0144929X.2019.1656779

Hair, Joe F., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. G. (2014). Partial least squares structural equation modeling (PLS-SEM): An emerging tool in business research. *European Business Review*, *26*(2), 106–121. https://doi.org/10.1108/EBR-10-2013-0128

Hair, Joseph F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, *31*(1), 2–24. https://doi.org/10.1108/EBR-11-2018-0203

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 2125–2126. https://doi.org/10.1145/2939672.2945386

Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology*, *15*(2), 99–107. https://doi.org/10.1007/s10676-012-9301-2

Hofeditz, L., Ehnis, C., Bunker, D., Brachten, F., & Stieglitz, S. (2019). Meaningful use of social bots? Possible applications in crisis communication during disasters. *27th European Conference on Information Systems (ECIS)*.

Hofeditz, L., Mirbabaie, M., Erle, L., Knoßalla, E., & Timm, L. (2022). Automating Crisis Communication in Public Institutions – Towards Ethical Conversational Agents That Support Trust Management. *Proceedings of the 17th International Conference on Wirtschaftsinformatik*.

Holtgraves, T., & Han, T. L. (2007). A procedure for studying online conversational processing using a chat bot. *Behavior Research Methods*, *39*(1), 156–163. https://doi.org/10.3758/BF03192855

Independent High-Level expert Group on Artifical Intelligence. (2019). *Ethics Guidelines for Trustworthy AI: Set up by the European Commission*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

James, R., & Boyles, M. (2017). Philosophical Signposts for Artificial Moral Agent Frameworks. *Suri*, *6*(2), 92–109. https://philpapers.org/rec/BOYPSF

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. Le, Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., & Choi, Y. (2021). *Can Machines Learn Morality? The Delphi Experiment*. http://arxiv.org/abs/2110.07574

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johannsen, F., Leist, S., Konadl, D., Basche, M., & de Heselle, B. (2018). Comparison of Commercial Chatbot solutions for Supporting Customer Interaction. *Twenty-Sixth European Conference on Information Systems*, 1–17.

Johnson, D. G., & Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology*, *7*(2), 99–107. https://doi.org/10.1007/s10676-005-4585-0

Kawaguchi, K. (2021). When will workers follow an algorithm? A field experiment with a retail business. *Management Science*, *67*(3), 1670–1695. https://doi.org/10.1287/mnsc.2020.3599

Khosrawi-Rad, B., Rinn, H., Schlimbach, R., Gebbing, P., Yang, X., Lattemann, C., Markgraf, D., & Robra-Bissantz, S. (2022). Conversational Agents in Education – A Systematic Literature Review. *European Conference on Information Systems*. https://aisel.aisnet.org/ecis2022_rp/18

Kington, T. (2022). *ChatGPT bot tricked into giving bomb-making instructions, say developers*. The Times. https://www.thetimes.co.uk/article/chatgpt-bot-tricked-into-giving-bomb-making-instructions-say-developers-rvktrxqb5

Kock, N., & Mayfield, M. (2015). PLS-based SEM Algorithms : The good neighbor assumption , collinearity, and nonlinearity. *Information Management and Business Review*, *7*(2), 113–130.

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, *00*(00), 1–22. https://doi.org/10.1080/0960085X.2021.1927212

Krämer, N. C., Lucas, G., Schmitt, L., & Gratch, J. (2018). Social snacking with a virtual agent – On the interrelation of need to belong and effects of social responsiveness when interacting with artificial entities. *International Journal of Human Computer Studies*, *109*, 112–121. https://doi.org/10.1016/j.ijhcs.2017.09.001

Kraus, D., Reibenspiess, V., & Eckhardt, A. (2019). *How Voice Can Change Customer Satisfaction : A Comparative Analysis between E-Commerce and Voice Commerce*. 1868–1879.

Krovi, R. (1993). Identifying the causes of resistance to IS implementation - A change theory perspective. *Information & Management*, *25*, 327–335.

Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, *13*(1), 4569. https://doi.org/10.1038/s41598-023-31341-0

Kumar, P., Dwivedi, Y. K., & Anand, A. (2021). Responsible Artificial Intelligence (AI) for Value Formation and Market Performance in Healthcare: the Mediating Role of Patient's Cognitive Engagement. *Information Systems Frontiers*, *25*, 2197–2220. https://doi.org/10.1007/s10796-021-10136-6

Lankton, N. K., Harrison Mcknight, D., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, *16*(10), 880–918. https://doi.org/10.17705/1jais.00411

Larsen, K. R., Hovorka, D. S., Dennis, A. R., & West, J. D. (2019). "Understanding the Elephant: The Discourse Approach to Boundary Identification and Corpus Construction for Theory Review Articles." *Journal of the Association for Information Systems*, *20*, 887–927. https://doi.org/10.17705/1jais.00556

Laumer, S., & Eckard, A. (2010). Why do People Reject Technologies?–Towards a Unified Model of Resistance to IT-Induced Organizational Change. *Diffusion Interest Group in Information Technology (DIGIT) Workshop*. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1013&context=digit2010&sei-redir=1

Laurim, V., Arpaci, S., Prommegger, B., & Krcmar, H. (2021). Computer, whom should I hire? - Acceptance criteria for artificial intelligence in the recruitment process. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2020-Janua*, 5495–5504. https://doi.org/10.24251/hicss.2021.668

Lee, J.-E. R., & Nass, C. I. (2010). Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In *Trust and technology in a ubiquitous modern environment:*

Theoretical and methodological perspectives (pp. 1–15). IGI Global. https://doi.org/10.4018/978-1-61520-901-9.ch001

Lee, S. K., Kavya, P., & Lasser, S. C. (2021). Social interactions and relationships with an intelligent virtual agent. *International Journal of Human Computer Studies*, *150*(May 2020), 102608. https://doi.org/10.1016/j.ijhcs.2021.102608

Lewicki, R. J., & Bunker, B. B. (1996). Developing and maintaining trust in work relationships. In *Trust in Organizations: Frontiers of Theory and Research* (pp. 114–139). SAGE Publications, Inc. https://doi.org/10.4135/9781452243610.n7

Lewicki, R. J., Stevenson, M. A., & Bunker, B. B. (1997). *The three components of interpersonal trust: instrument development and differences across relationships*. Max M. Fisher College of Business, Ohio State University.

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human Computer Studies*, *77*, 23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001

Liddell, H. G. (1889). *An Intermediate Greek-English Lexicon* (A. B. Company (ed.)). Harper & Brothers.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*(April 2018), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Luco, A. (2014). The Definition of Morality. *Social Theory and Practice*, *40*(3), 361–387. https://doi.org/10.5840/soctheorpract201440324

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants. *Business & Information Systems Engineering*, *61*(4), 535–544. https://doi.org/10.1007/s12599-019-00600-8

Marcum, J. A. (2012). Virtues and Vices. *Philosophy and Medicine*, *114*, 59–105. https://doi.org/10.1007/978-94-007-2706-9_3

Martin, K. (2019). Designing ethical algorithms. *MIS Quarterly Executive*, *18*(2), 129–142. https://doi.org/10.17705/2msqe.00012

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709. https://doi.org/10.2307/258792

Mayring, P. (2014). Qualitative content analysis: theoretical foundation, basic procedures and software solution. In *gesis: Social Science Open Access Repository (SSOAR)*. gesis. https://doi.org/https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, *2*(2). https://doi.org/10.1145/1985347.1985353

Meinhardt-Injac, B., Daum, M. M., Meinhardt, G., & Persike, M. (2018). The two-systems account of theory of mind: Testing the links to social-perceptual and cognitive abilities. *Frontiers in Human Neuroscience*, *12*(January), 1–12. https://doi.org/10.3389/fnhum.2018.00025

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, *39*(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Meyer, S., Mandl, S., Gesmann-Nuissl, D., & Strobel, A. (2023). Responsibility in Hybrid Societies: concepts and terms. *AI and Ethics*, *3*(1), 25–48. https://doi.org/10.1007/s43681-022-00184-2

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information and Management*, *58*(3), 103434. https://doi.org/10.1016/j.im.2021.103434

Mingers, J., Mutch, A., & Willcocks, L. (2013). Critical Realism in Information Systems Research. *MIS Quarterly*, *37*(3), 795–802.

Mingers, J., & Walsham, G. (2010). Toward Ethical Information Systems: The Contribution of Discourse Ethics. *MIS Quarterly*, *34*(4), 833–854.

Mirbabaie, M., Hofeditz, L., Frick, N. R. J., & Stieglitz, S. (2022). Artificial intelligence in hospitals: providing a status quo of ethical considerations in academia to guide future research. *AI & Society*, *37*, 1361–1382. https://doi.org/10.1007/s00146-021-01239-4

Mirbabaie, M., Stieglitz, S., Brünker, F., Hofeditz, L., Ross, B., & Frick, N. R. J. (2021). Understanding Collaboration with Virtual Assistants – The Role of Social Identity and the Extended Self. *Business and Information Systems Engineering*, *63*(1), 21–37. https://doi.org/10.1007/s12599-020-00672-x

Mirbabaie, M., Stieglitz, S., & Frick, N. R. J. (2021). Hybrid intelligence in hospitals: towards a research agenda for collaboration. *Electronic Markets*. https://doi.org/10.1007/s12525-021-00457-4

Miroshnichenko, A. (2018). AI to bypass creativity. Will robots replace journalists? (The answer is "yes"). *Information (Switzerland)*, *9*(7), 1–20. https://doi.org/10.3390/info9070183

Mitchell, P. (1997). Introduction to theory of mind: Children, autism and apes. In *Introduction to theory of mind: Children, autism and apes.* Edward Arnold Publishers.

Mittelstadt, B. (2016). Auditing for transparency in content personalization systems. *International Journal of Communication*, *10*(October), 4991–5002.

Möhlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R. W. (2021). Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control. *MIS Quarterly*, *45*(4), 1999–2022. https://doi.org/10.25300/misq/2021/15333

Momen, A., De Visser, E., Walliser, J., De Visser, E. J., Wolsten, K., Cooley, K., & Tossell, C. C. (2023). Trusting the Moral Judgments of a Robot: Perceived Moral Competence and Humanlikeness of a GPT-3 Enabled AI. *Proceedings of the 56th Hawaii International Conference on System Sciences*.

Moradi, M., & Samwald, M. (2021). Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications*, *165*, 113941. https://doi.org/10.1016/j.eswa.2020.113941

Mumford, E., Flores-Saviaga, C., Savage, S., & Taraborelli, D. (2006). *Version of Record: https://www.sciencedirect.com/science/article/pii/S1071581917301271*. 317–342. https://doi.org/10.1145/2818052.2869106

Newswire. (2023). *Now There's a "Text to Image" AI Without Censorship*. Newswire. https://www.newswire.com/news/now-theres-a-text-to-image-ai-without-censorship-22143108

Nissen, A., Krampe, C., Kenning, P., & Schütte, R. (2019). Utilizing Mobile FNIRS to Investigate Neural Correlates of the TAM in ECommerce. *International Conference on Information Systems (ICIS)*, 1–9.

Nißen, M., Selimi, D., Janssen, A., Cardona, D. R., Breitner, M. H., Kowatsch, T., & von Wangenheim, F. (2022). See you soon again, chatbot? A design taxonomy to characterize user-chatbot relationships with different time horizons. *Computers in Human Behavior*, *127*(September 2021). https://doi.org/10.1016/j.chb.2021.107043

Noy, S., & Zhang, W. (2023). *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence* (Issue 1745302). https://doi.org/http://dx.doi.org/10.2139/ssrn.4375283

Nyholm, S. (2018). Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*, *24*(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishers.

Ochmann, J., & Laumer, S. (2019). Fairness as a Determinant of AI Adoption in Recruiting: An Interview-based Study. *DIGIT 2019 Proceedings*. https://aisel.aisnet.org/digit2019/16

Ochmann, J., Zilker, S., Michels, L., Tiefenbeck, V., & Laumer, S. (2021). The influence of algorithm aversion and anthropomorphic agent design on the acceptance of AI-based job recommendations. *International Conference on Information Systems, ICIS 2020*, 0–17.

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2020). The present and future use of functional near-infrared spectroscopy (Fnirs) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1464*(1), 5–29. https://doi.org/10.1111/nyas.13948

Porra, J., Lacity, M., & Parks, M. S. (2020). "Can Computer Based Human-Likeness Endanger Humanness?" – A Philosophical and Ethical Perspective on Digital Assistants Expressing Feelings They Can't Have". *Information Systems Frontiers*, *22*(3), 533–547. https://doi.org/10.1007/s10796-019-09969-z

Prakash, A. V., & Das, S. (2021). Medical practitioner's adoption of intelligent clinical diagnostic decision support systems: A mixed-methods study. *Information and Management*, *58*(7), 103524. https://doi.org/10.1016/j.im.2021.103524

Rai, A., Constantinides, P., & Sarker, S. (2019). Next-Generation Digital Platforms: Toward Human-AI Hybrids. *MIS Quarterly*, *43*(1), iii–ix.

Reddy, S., Fox, J., & Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine*, *112*(1), 22–28. https://doi.org/10.1177/0141076818815510

Reis, L., Maier, C., & Weitzel, T. (2022). Mixed-Methods in Information Systems Research: Status Quo, Core Concepts, and Future Research Implications. *Communications of the Association for Information Systems*, *51*(1), 95–119. https://doi.org/10.17705/1CAIS.05106

Renier, L. A., Schmid Mast, M., & Bekbergenova, A. (2021). To err is human, not algorithmic – Robust reactions to erring algorithms. *Computers in Human Behavior*, *124*. https://doi.org/10.1016/j.chb.2021.106879

Robinson, B. (2020). Towards an ontology and ethics of virtual influencers. *Australasian Journal of Information Systems*, *24*, 1–8. https://doi.org/10.3127/AJIS.V24I0.2807

Rousseau, D. M., Sitkin, S., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross discipline view of trust. *Academy of Management Review*, *23*(3), 393–404.

Russel, S., & Norvig, P. (2012). Artificial intelligence—a modern approach 3rd Edition. In *The Knowledge Engineering Review*. https://doi.org/10.1017/S0269888900007724

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). *An integrative model of organizational trust: Past, present, and future*. Academy of Management Briarcliff Manor, NY 10510.

Schulze, L., Trenz, M., & Cai, Z. (2022). Algorithmic unfairness on digital labor platforms: How algorithmic management practices disadvantage workers.

International Conference on Information Systems (ICIS).

Schulze, L., Trenz, M., Cai, Z., & Tan, C.-W. (2023). Fairness in Algorithmic Management: How Practices Promote Fairness and Redress Unfairness on Digital Labor Platforms. *Hawaii International Conference on System Sciences*.

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G. J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information and Management*, *57*(2), 103174. https://doi.org/10.1016/j.im.2019.103174

Seele, P., & Schultz, M. D. (2022). From Greenwashing to Machinewashing: A Model and Future Directions Derived from Reasoning by Analogy. *Journal of Business Ethics*, *178*(4), 1063–1089. https://doi.org/10.1007/s10551-022-05054-9

Seidel, S., & Watson, R. T. (2020). Integrating explanatory/predictive and prescriptive science in information systems research. *Communications of the Association for Information Systems*, *47*, 284–314. https://doi.org/10.17705/1CAIS.04714

Seymour, M., Yuan, L., Dennis, A. R., & Riemer, K. (2021). Have we crossed the uncanny valley? Understanding affinity, trustworthiness, and preference for realistic digital humans in immersive environments. *Journal of the Association for Information Systems*, *22*(3), 591–617. https://doi.org/10.17705/1jais.00674

Seymour, M., Yuan, L., Dennis, A. R., & Riemer, K. (2020). Facing the artificial: Understanding affinity, trustworthiness, and preference for more realistic digital humans. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 4673–4683. https://doi.org/10.24251/hicss.2020.574

Shawar, B. A., & Atwell, E. S. (2005). Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics*, *10*(4), 489–516. https://doi.org/10.1075/ijcl.10.4.06sha

Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, *10*(4). https://doi.org/10.1145/3419764

Short, C. E., & Short, J. C. (2023). The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation. *Journal of Business Venturing Insights*, *19*, e00388. https://doi.org/https://doi.org/10.1016/j.jbvi.2023.e00388

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics. *Journal of Database Management*, *31*(2), 74–87. https://doi.org/10.4018/jdm.2020040105

Simpson, T. W. (2012). What is trust? *Pacific Philosophical Quarterly*, *93*(4), 550–569. https://doi.org/10.1111/j.1468-0114.2012.01438.x

Simpson, T. W. (2013). Trustworthiness and Moral Character. *Ethical Theory and Moral Practice*, *16*(3), 543–557. https://doi.org/10.1007/s10677-012-9373-4

Skerker, M., Purves, D., & Jenkins, R. (2020). Autonomous weapons systems and the moral equality of combatants. *Ethics and Information Technology*, *22*(3), 197–209. https://doi.org/10.1007/s10676-020-09528-0

Sowa, K., Przegalinska, A., & Ciechanowski, L. (2021). Cobots in knowledge work. *Journal of Business Research*, *125*(November 2020), 135–142. https://doi.org/10.1016/j.jbusres.2020.11.038

Spiekermann, S., Krasnova, H., Hinz, O., Baumann, A., Benlian, A., Gimpel, H., Heimbach, I., Köster, A., Maedche, A., Niehaves, B., Risius, M., & Trenz, M. (2022). Values and Ethics in Information Systems. *Business & Information Systems Engineering*. https://doi.org/10.1007/s12599-021-00734-8

Stahl, B. C. (2012). Morality, ethics, and reflection: A categorization of normative IS research. *Journal of the Association of Information Systems*, *13*(8), 636–656.

https://doi.org/10.17705/1jais.00304

Starke, G., De Clercq, E., & Elger, B. S. (2021). Towards a pragmatist dealing with algorithmic bias in medical machine learning. *Medicine, Health Care and Philosophy*, *24*(3), 341–349. https://doi.org/10.1007/s11019-021-10008-5

Stenseke, J. (2021). Artificial virtuous agents: from theory to machine implementation. *AI and Society*, *0123456789*. https://doi.org/10.1007/s00146-021-01325-7

Stieglitz, S., Brachten, F., & Kissmer, T. (2018). Defining bots in an enterprise context. *International Conference on Information Systems 2018, ICIS 2018*, *Munger 2017*, 1–9.

Stieglitz, S., Hofeditz, L., Brünker, F., Ehnis, C., Mirbabaie, M., & Ross, B. (2022). Design principles for conversational agents to support Emergency Management Agencies. *International Journal of Information Management*, *63*, 102469. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2021.102469

Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. *Nature*, *614*(7947), 214–216. https://doi.org/10.1038/d41586-023-00340-6

Strack, M., & Gennerich, C. (2007). Erfahrung mit Forsyths 'Ethic Position Questionnaire? (EPQ): Bedeutungsunabhängigkeit von Idealismus und Realismus oder Akquieszens und Biplorarität? *Berichte Aus Der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral", Nr. 167, ISSN 1430-1148*.

Suárez-Gonzalo, S., Mas-Manchón, L., & Guerrero-Solé, F. (2019). Tay is You. The attribution of responsibility in the algorithmic culture. *Observatorio*, *13*(2), 1–14. https://doi.org/10.15847/obsOBS13220191432

Suen, H. Y., & Hung, K. E. (2023). Building trust in automatic video interviews using various AI interfaces: Tangibility, immediacy, and transparency. *Computers in Human Behavior*, *143*(December 2022), 107713. https://doi.org/10.1016/j.chb.2023.107713

Tajfel, H., & Turner, J. C. (2004). The social identity theory of intergroup behavior. In *Political psychology* (pp. 276–293). Psychology Press.

Tavanapour, N., Poser, M., & Bittner, E. A. C. (2019). Supporting the idea generation process in citizen participation - Toward an interactive system with a conversational agent as facilitator. *27th European Conference on Information Systems*, 0–17.

Teodorescu, M. H. M., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of fairness in automation require a deeper understanding of human–ml augmentation. *Management Information Systems Quarterly*, *45*(3), 1483–1499. https://doi.org/10.25300/MISQ/2021/16535

Tigard, D. W. (2021). Responsible AI and moral responsibility: a common appreciation. *AI and Ethics*, *1*(2), 113–117. https://doi.org/10.1007/s43681-020-00009-0

van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*(February), 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

Verbeek, P.-P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press.

Vinanzi, S., Patacchiola, M., Chella, A., & Cangelosi, A. (2019). Would a robot trust you? Developmental robotics model of trust and theory of mind. *CEUR Workshop Proceedings*, *2418*, 74.

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for*

Information Systems, *37*, 205–224. https://doi.org/10.17705/1CAIS.03709

Wallace, E., Buil, I., & de Chernatony, L. (2020). 'Consuming Good' on Social Media: What Can Conspicuous Virtue Signalling on Facebook Tell Us About Prosocial and Unethical Intentions? *Journal of Business Ethics*, *162*(3), 577–592. https://doi.org/10.1007/s10551-018-3999-7

Wang, W., & Benbasat, I. (2008). Attributions of trust in decision support technologies: A study of recommendation agents for e-commerce. *Journal of Management Information Systems*, *24*(4), 249–273. https://doi.org/10.2753/MIS0742-1222240410

Webster, J., & Watson, T. R. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2).

Wischmeyer, T., & Rademacher, T. (2019). Regulating artificial intelligence. In *Regulating Artificial Intelligence*. Springer. https://doi.org/10.1007/978-3-030-32361-5

Yang, R., & Wibowo, S. (2022). User trust in artificial intelligence: A comprehensive conceptual framework. *Electronic Markets 2022 32:4*, *32*(4), 2053–2077. https://doi.org/10.1007/S12525-022-00592-6

You, S., Yang, C. L., & Li, X. (2022). Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *Journal of Management Information Systems*, *39*(2), 336–365. https://doi.org/10.1080/07421222.2022.2063553

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). *Exploring AI Ethics of ChatGPT: A Diagnostic Analysis*. http://arxiv.org/abs/2301.12867

Appendix

Publications

Paper 1: Applying XAI to an AI-based System for Candidate Management to Mitigate Bias and Discrimination in Hiring

Type (Ranking, Impact Factor)	Journal article (B, 8.5)
Status	Published
Rights and permissions	Open access
Authors	Hofeditz, L., Clausen, S., Rieß, A., Mirbabaie, M., & Stieglitz, S.
Year	2022
Outlet	Electronic Markets (ELMA)
Permalink / DOI	https://doi.org/10.1007/s12525-022-00600-9
Full citation	Hofeditz, L., Clausen, S., Rieß, A., Mirbabaie, M., & Stieglitz, S. (2022). Applying XAI to an AI-based System for Candidate Management to Mitigate Bias and Discrimination in Hiring. *Electronic Markets*, 32, 2207–2233. https://doi.org/10.1007/s12525-022-00600-9.

Electronic Markets (2022) 32:2207–2233
https://doi.org/10.1007/s12525-022-00600-9

RESEARCH PAPER

Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring

Lennart Hofeditz[1] · Sünje Clausen[1] · Alexander Rieß[1] · Milad Mirbabaie[2] · Stefan Stieglitz[1]

Received: 20 May 2022 / Accepted: 30 September 2022 / Published online: 20 December 2022

Abstract

Assuming that potential biases of Artificial Intelligence (AI)-based systems can be identified and controlled for (e.g., by providing high quality training data), employing such systems to augment human resource (HR)-decision makers in candidate selection provides an opportunity to make selection processes more objective. However, as the final hiring decision is likely to remain with humans, prevalent human biases could still cause discrimination. This work investigates the impact of an AI-based system's candidate recommendations on humans' hiring decisions and how this relation could be moderated by an Explainable AI (XAI) approach. We used a self-developed platform and conducted an online experiment with 194 participants. Our quantitative and qualitative findings suggest that the recommendations of an AI-based system can reduce discrimination against older and female candidates but appear to cause fewer selections of foreign-race candidates. Contrary to our expectations, the same XAI approach moderated these effects differently depending on the context.

Keywords Explainable AI · Hiring · Bias · Discrimination · Ethics

JEL Classification O30

Introduction

At 99% of Fortune 500 companies, job applications are first evaluated by an applicant tracking system instead of a human being (Hu, 2019). These systems are often based on artificial intelligence (AI) and allow human resource (HR) professionals to cope with large amounts of applicant data, the pressure to give timely responses to candidates, and limited resources for finding the best talent (Mujtaba & Mahapatra, 2019; Raghavan et al., 2020). While a universally accepted definition does not exist, AI has recently been defined as "the frontier of computational advancements that references human intelligence in addressing ever more complex decision-making problems" (Berente et al., 2021, p. 1,435). Thus, AI refers to machines performing a spectrum of cognitive tasks and intelligent behavior patterns commonly associated with human intelligence (Russell & Norvig, 2016). AI comprises a variety of methods, such as machine learning (ML) and rule-based symbolic logic, which differ in their complexity and suitability for different tasks (Rouse, 2020). To date, strong AI that is akin to human intelligence does not exist. The present research focuses on a type of so-called weak AI that simulates intelligent behavior in a certain area, specifically on using ML to identify suitable candidates among job applicants (Russell & Norvig, 2016).

Importantly, AI-based systems also promise to combat the pressing problem of discrimination in hiring (Quillian et al., 2017; Sánchez-Monedero et al., 2020; Zschirnt & Ruedin, 2016) by basing decisions solely on skillsets

Responsible Editor: Babak Abedin

✉ Lennart Hofeditz
lennart.hofeditz@uni-due.de

Sünje Clausen
suenje.clausen@uni-due.de

Alexander Rieß
alexander.riess@uni-due.de

Milad Mirbabaie
milad.mirbabaie@uni-paderborn.de

Stefan Stieglitz
stefan.stieglitz@uni-due.de

[1] Universität Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany

[2] Paderborn University, Warburger Str. 100, 33098 Paderborn, Germany

and criteria related to job requirements rather than additional information such as demographic criteria to reduce the impact of human biases (Li et al., 2021; Ochmann & Laumer, 2019). However, this process can still be challenging, as some criteria, such as social skills, are difficult to measure using an AI-based system, and it is often difficult for humans to comprehend a system's output. While previous literature and news media have raised concerns about potential biases in AI-based systems (Barocas & Selbst, 2016; Raghavan et al., 2020), such as the preference for male candidates in Amazon's recruitment system (Dastin, 2018), machines themselves cannot be moral or immoral. Instead, biases in the historical data used to train an AI-based system lead to biased results, referred to as "garbage in, garbage out" (Barocas & Selbst, 2016). Discrimination by AI can also result from algorithms and presentations (Kulshrestha et al., 2019; Wijnhoven & van Haren, 2021). However, AI-based systems make such biases visible and controllable and thus can not only lead to more successful hires and lower costs but also reduce discrimination and facilitate diversity in hiring (e.g., Houser, 2019; Li et al., 2021). Nonetheless, attempts by organizations as large as Amazon to automate the hiring process have failed, which indicates that humans are still needed as final decision makers (Dastin, 2018).

Yet, AI-based systems are not likely to entirely replace humans in hiring soon but rather to augment human decision-making (Ebel et al., 2021). Augmentation refers to application scenarios of AI in organizations in which "humans collaborate closely with machines to perform a task" (Raisch & Krakowski, 2021, p. 193). Therefore, augmentation can take different forms depending on, for example, whether the AI or the human agent makes the final decision (Teodorescu et al., 2021). Here, we investigate a type of augmentation where the human is the locus of the decision, that is, where the human is the final decision maker. Thus, in this scenario, the AI will not take over the task of hiring but collaborate with the human to identify suitable candidates (Raisch & Krakowski, 2021). The final decision on whom to hire remains with the human, which introduces potential barriers to realizing the potential of AI-based systems in hiring. Some people prefer to retain decision-making power and tend to be averse to the decisions and predictions of AI-based systems and similar algorithms (Berger et al., 2021; Dietvorst et al., 2015; Jussupow et al., 2020; Ochmann et al., 2021). This phenomenon occurs even if the algorithm's predictions are better than those of humans. High self-confidence in particular has been shown to reduce the acceptance of advice from an AI-based system (Chong et al., 2022). One reason might be that the origin of recommendations made by an AI-based system is often incomprehensible (Adadi & Berrada, 2018), which makes it difficult for people to trust the underlying technology (Zhu et al., 2018). This could lead to scenarios in which an AI-based system recommends an objectively better-qualified applicant, but the human chooses another applicant nonetheless. Thus, the final human decision could still systematically disadvantage racial minorities, older and very young applicants, and female applicants (Baert, 2018). Thus, to encourage humans to follow the recommendations of AI-based systems, additional mechanisms are needed. Accordingly, we formulated the following research question (RQ):

> **RQ1:** *Given comparable candidate qualifications, how can an AI-based system's recommendation reduce discrimination (based on the sensitive attributes race, age, gender) in hiring decisions?*

The field of explainable AI (XAI) seeks to provide better insight into how and why an AI-based system operates the way it does (Adadi & Berrada, 2018). Barredo Arrieta et al. (2020) defined XAI as follows: "Given a certain audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand" (p. 6). XAI refers to a variety of approaches (e.g., reverse engineering) to overcome the opaque nature of some types of AI-based systems, such as deep neural networks (Guidotti et al., 2018; Meske et al., 2022). Thereby, different XAI approaches serve different purposes and should be tailored to the target audience (Barredo Arrieta et al., 2020; Meske et al., 2022). As the target audience for the system investigated in this study comprises individuals managing applicants in hiring, we adopt a type of XAI that provides users with high-level insights into how the AI-based system weighs (sensitive) candidate attributes to derive candidate recommendations. Previous research has attempted to design XAI in a more human-centered way by testing the effect of providing contextual domain knowledge, which was found to be an influencing factor on trust of and reliance on AI-based systems (Dikmen & Burns, 2022). XAI can increase users' trust in an AI-based system's recommendations, their knowledge about the system, and the decision task (Barredo Arrieta et al., 2020; Meske et al., 2022). As the implementation of XAI has been shown to increase trust in AI (Meske & Bunde, 2020), XAI could improve user confidence in candidate recommendations by an AI-based system (Gunning et al., 2019; Hoffman et al., 2018). Therefore, we state a second research question:

> **RQ2:** *What is the influence of explainable AI on decision-making in the context of an AI-based system's recommendation for hiring decisions?*

Implementing XAI in AI-based systems for candidate recommendations might increase human acceptance of these recommendations and, thus, reduce discrimination in hiring. However, little empirical research, which also shows contradictory results in terms of the effect of adding XAI and transparency, is available to date (Hofeditz et al., 2021; Shin, 2021). Previous research has indicated that the human's role is not sufficiently studied in existing explainability approaches (Adadi & Berrada, 2018). Therefore, it is also important to identify and understand the reasons for user hiring decisions on XAI-based candidate management platforms. Therefore, we pose a third research question:

> **RQ3:** *What are users' reasons for selecting applicants on an XAI-based candidate management platform?*

To address these research questions, we developed an interactive, functional prototype that simulated an AI-based system for candidate management and evaluated the impact of XAI and AI recommendations on the selection of typically disadvantaged individuals (2 × 2 between-subjects design, N = 194). As discrimination can differ between countries, we focused on a specific country and chose a German context for our study.

The remainder of the paper is structured as follows: First, we review relevant literature on AI-based systems in hiring, biases, and discrimination in hiring processes, and XAI. In the methods section, we describe the sample, the development of the prototypical AI-based system for candidate management, and the employed questionnaires. We then present quantitative and qualitative insights from our study and discuss them in the context of the relevant literature. The paper concludes with limitations, opportunities for future research, and a short summary of the main findings.

Related work

AI-based systems in hiring

As previously mentioned, AI is the frontier of computational advancements and refers to machines performing a spectrum of cognitive tasks commonly associated with human intelligence, such as complex decision-making (Berente et al., 2021; Russell & Norvig, 2016). In hiring, the use of AI-based systems has been on the rise in recent years (Black & van Esch, 2020; Raghavan et al., 2020), and organizations already use various software solutions in practice for hiring workers (Li et al., 2021; Raghavan et al., 2020; Sánchez-Monedero et al., 2020). While there has been limited research on the topic (Pan et al., 2021), existing literature suggests that AI-based systems can add great value to data-intensive and time-consuming processes in hiring, such as sourcing, screening, and the assessment of potential candidates (Black & van Esch, 2020; Li et al., 2021). Although Kuncel et al. (2014) stated that humans are good at defining job characteristics and assessing candidates in job interviews, in an analysis of 17 studies on candidate screening, they found that algorithms outperform human decision-making (measured in terms of the number of above-average performing employees recruited) by more than 25% if a large number of candidates must be screened. In addition to efficiency gains, AI-based systems also promise to reduce discrimination in hiring. Li et al. (2021) interviewed HR professionals who already used AI-based systems. Their findings suggest that the automation of hiring processes reduces opportunities for introducing biases and discrimination in hiring decisions and increases the diversity of hires (Li et al., 2021). Similarly, Ochmann and Laumer (2019) conducted expert interviews in HR management and suggested that AI can be used to highlight human biases and thus result in greater objectivity in decision-making (Ochmann & Laumer, 2019).

Black and van Esch (2020) presented several real-world examples of successful implementation of AI-based systems in organizations. For example, by introducing game-based assessments and video-based assessments, Unilever reduced the required time of HR professionals per application by 75% (Feloni, 2017). Typically, these systems do not replace but rather augment human decision-making, for example, by recommending the most suitable candidates for a position. Thus, the final hiring decision remains with the human, which poses the risk that human biases might still affect the selection of candidates.

Bias and discrimination in hiring

As AI-based systems in hiring are not expected to fully automate but instead augment decision-making, human biases might still allow discriminatory behavior. Hiring remains an area where discrimination is most common (Sánchez-Monedero et al., 2020; Zschirnt & Ruedin, 2016). Discrimination can result from a number of psychological reasons and occurs especially in contexts with limited or missing information (Fiske et al., 1991; Foschi et al., 1994; Tosi & Einbender, 1985), as is the case in hiring. When decision makers must make decisions based on limited information, they tend to rely more on a group's average performance to judge individuals (Guryan & Charles, 2013), and the likelihood of stereotyping increases (Fiske et al., 1991; Tosi & Einbender, 1985). In addition, ambiguous information allows room for interpretation by the human decision maker, which in turn may reinforce discrimination, as the decision is then more likely to be made based on stereotypes due to the cognitive models activated in these situations (Fiske et al., 1991). Difficulty documenting or tracking decision-making in hiring can increase discrimination as unethical behavior (Petersen & Saporta, 2004). A lack of documentation often implies that discrimination cannot be proven (Sabeg &

Me´haignerie, 2006), and thus, decision makers do not face negative consequences for unethical behavior. To mitigate unethical human behavior, previous research has suggested applying AI-based systems in hiring (Hofeditz et al., 2022a, 2022b; Sühr et al., 2021), as AI is already being used by some organizations to perform the preselection of applicants (Laurim et al., 2021). However, in practice, these systems are not in charge of making the final decision (without a human decision maker). What AI-based systems usually do is provide recommendations to augment human decision-making in organizations that target in a certain direction. XAI in combination with the provision of domain knowledge can help increase trust in AI-based systems (Dikmen & Burns, 2022). With AI-based recommendations, we assume that XAI both challenges human assumptions and augments human decision-making by providing information that the human otherwise would not be aware of.

On the one hand, an AI-based system might encourage reflection on the (objective) reasons for selecting a candidate, especially if the candidate preferred by the human and the recommendation of the AI-based system differ (Ochmann & Laumer, 2019). On the other hand, it is important that the AI-based system's recommendations not be discriminatory by design or based on certain data. Previous research has already focused on approaches to how AI-based systems can be applied without causing discrimination in hiring by, for example, avoiding biases in historical data (van Giffen et al., 2022). In this study, we therefore assume that AI-based systems in hiring are blind to historical demographic characteristics and increasingly provide recommendations based solely on objective criteria. Rieskamp et al. (2023) summarized different approaches that aim to mitigate AI-based systems' discrimination by building on pre-process, in-process, post-process, and feature-selection approaches. Using a pre-process approach, historical data can be normalized for the training of the algorithm. Thus, if it can be assumed that AI-based systems in hiring increasingly embrace diversity, it is important to focus on human decision makers as the origin of discrimination.

Discrimination in hiring is highly relevant and frequently discussed in the literature (Akinlade et al., 2020; Baert et al., 2017; Neumark, 2018; Quillian et al., 2017, 2019; Zschirnt & Ruedin, 2016). A recent meta-analysis by Zschirnt and Ruedin (2016) found that candidates from minority groups must send out approximately 50% more applications to be invited for a job interview. Ameri et al. (2018) showed that applicants who indicated a disability that would not affect job performance received 26% less feedback than those not indicating a disability. Baert (2018) comprehensively evaluated empirical studies on discrimination in hiring from 2005 to 2016 and identified race, gender, age, religion, sexual orientation, disability, and physical appearance as reasons for discrimination that are sufficiently supported by the literature; age, gender, and race are the most frequently mentioned reasons for discrimination in the literature (Baert, 2018). We also found that these three forms of discrimination are the most common in online hiring, which made them the most suitable for our study (see Table 6 in Appendix 1 for an overview of reasons for discrimination).

An extensive amount of literature suggests addressing the issue of racial discrimination in hiring (Lancee, 2021; Quillian et al., 2017, 2019; Zschirnt & Ruedin, 2016). For example, Lancee (2021) found in a cross-national study that ethnic minorities have significantly lower chances of being hired. Quillian et al. (2017) suggested that the rate of discrimination in hiring against African Americans has not decreased over the past 25 years in the United States. Thus, race-based discrimination in hiring is among the most important cases needing to be considered, and action must be taken to ensure that it does not continue.

Victims of discrimination can differ among cultures and countries. Quillian et al. (2019) were able to show that racial discrimination in Germany occurs mostly against Turkish candidates. As we focused on the German context in this study, we chose job applicants with a Turkish name to test for race-based discrimination.

Building on the literature suggesting that AI-based systems can reduce discrimination in hiring, we hypothesize the following:

H1: *Recommending foreign-race candidates in an AI-based system for candidate management leads to a higher rate of foreign-race candidate selection.*

Age-based discrimination is also one of the most relevant issues in hiring (Abrams et al., 2016; Baert, 2018; Lössbroek et al., 2021; Neumark et al., 2017; Zaniboni et al., 2019). Although this form of discrimination can affect both "too young" and "too old" candidates, current literature states that older applicants tend to have worse job application chances than younger applicants (Lössbroek et al., 2021; Neumark, 2021; Zaniboni et al., 2019). Reasons for this can be stereotypical perceptions of older candidates, such as poorer trainability (Richardson et al., 2013). Richardson et al. (2013) also found that applicants in the age group of 42 to 48 years are preferred and hired more frequently than older or younger applicants. There is also evidence in the literature that little work experience is more often a stereotypical perception of younger candidates (Baert et al., 2017). Therefore, it was important that the control group of candidates in this study be neither too old nor too young. Therefore, this study compared applicants who were younger (33–39 years old) or older (51–57 years old). A possible approach to reducing discrimination against older candidates in hiring could be the use of an AI-based candidate recommendation system, as previous research has examined their potential in recruiting (Mehrotra & Celis, 2021). Therefore, the following hypothesis is proposed:

H2: *Recommending older candidates in an AI-based system for candidate management leads to a higher rate of older candidate selection.*

Previous literature clearly shows that men are consistently preferred over women in application processes (Baert, 2018; Carlsson & Sinclair, 2018; Kübler et al., 2018). The literature suggests that discrimination in hiring processes has led to and reinforces this gender inequity (Petersen & Togstad, 2006), and discrimination is often based on stereotypes, such as lower productivity of female applicants (González et al., 2019). Furthermore, Correll et al. (2007) found that women are penalized for motherhood in hiring due to various factors, such as being family-oriented. Another study found that female recruiters attributed more work experience to male applicants' resumes than equal female applicants' resumes (Cole et al., 2004), suggesting that even female recruiters discriminate against female applicants. In addition to male and female applicants, other types of gender experience discrimination in hiring (Davidson, 2016). However, this study was conducted in a yet unexplored field of research. For simplicity, the binary gender system was used to compare male and female applicants in this study. A possible approach to reducing discrimination of female candidates in hiring might be the use of an AI-based system for candidate recommendations, as gender is also a source of discrimination that has already been examined in the context of AI-based systems in previous studies (Fernández-Martínez & Fernández, 2020; Köchling et al., 2021). Therefore, we hypothesize:

H3: *Recommending female candidates in an AI-based system for candidate management leads to a higher rate of female candidate selection.*

We use the term "sensitive attributes" to describe characteristics of candidates who are of older age, foreign race, or female and consider these attributes in the context of an AI-based system for candidate management.

Explainable AI and its role in decision-making

One challenge in working with AI-based systems is that their results cannot always be easily explained or tracked (Dwivedi et al., 2021), and artificial and deep neural networks in particular have been described as a black box (Adadi & Berrada, 2018). This is a problem, especially for high-stakes or sensitive decisions such as hiring, as it is often not possible to explain why a system produced a certain result (Gunning et al., 2019; Hepenstal & McNeish, 2020; Sokol & Flach, 2020).

The general aim of implementing XAI is to disclose the behavior of the AI to users and make it comprehensible (Barredo Arrieta et al., 2020; Gunning et al., 2019). However, due to the relative newness and large quantity of XAI research, a standardized understanding and precise terminology regarding the term "XAI" and its applications are missing (Barredo Arrieta et al., 2020; Hussain et al., 2021; Meske et al., 2022).

Several recent surveys have provided an overview and categorization of technical XAI approaches (Adadi & Berrada, 2018; Gilpin et al., 2018; Guidotti et al., 2018). For example, Gilpin et al. (2018) focused on explaining deep neural architectures and propose a taxonomy consisting of three categories of XAI approaches that respectively: i) emulate the processing of the data, ii) explain the representation of data inside the network, or iii) are explanation-producing. Despite being less technically specific, the XAI type explored in this work falls most closely into the first category in that providing some form of justification for the input–output relation of the system may "build human trust in the system's accuracy and reasonableness" (Gilpin et al., 2018, p. 86).

However, XAI is not just a technical concept but a movement, initiative, or effort in response to transparency issues related to AI-based systems (Adadi & Berrada, 2018). Similarly, Barredo Arrieta et al. (2020) stated that "any means to reduce the complexity of the model or simplify its outputs should be considered as an XAI approach" (p. 6). In selecting an appropriate XAI approach, Meske et al. (2022) argued that there are different objectives for XAI and the stakeholders for whom XAI is relevant. In this study, we focused on *users* of AI-based systems, for whom XAI can increase trust in the system's recommendation and allow them to compare the system's reasoning with their own. Furthermore, a main objective of XAI that we consider in this work is that users be able to learn from an AI-based system. Thus, in this research, we did not focus on a highly technical XAI approach (e.g., for explaining deep neural architectures as described by Gilpin et al., 2018) but provided users with a high-level explanation of how the AI-based system selects candidates and how it considers the sensitive attributes of the candidates. Thereby, the XAI can help users gain knowledge on diverse hiring selection decisions.

In XAI research, the term "transparency" often appears but is then not sufficiently differentiated. AI transparency and XAI have some overlap and are difficult to consider separately. Whereas AI transparency can be limited to the mere visibility of the deployment or use of an AI, XAI takes one step beyond this and aims to provide easily understandable and comprehensible explanations and derivations of the procedure and output of an AI-based system (Schmidt et al., 2020). Simplified, the relationship between XAI and

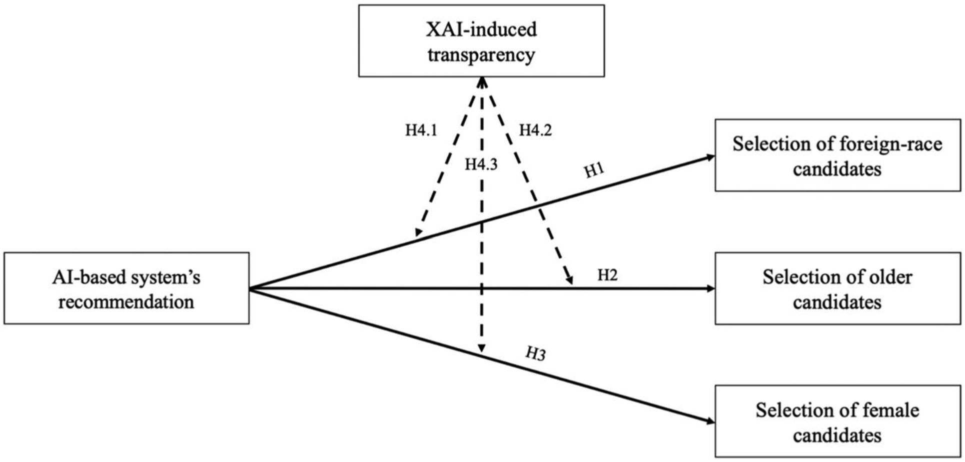

XAI-induced transparency
H4.1
H4.2
H4.3
H1
H2
H3
AI-based system's recommendation
Selection of foreign-race candidates
Selection of older candidates
Selection of female candidates

Fig. 1 Visualization of the proposed research model

transparency is that XAI is an approach or an effort made in response to a need for more transparency for stakeholders, such as decision makers, in the context of using AI-based systems (Adadi & Berra, 2018). However, in our literature research, we found that the distinction and relation between XAI and transparency is often not clearly addressed. With XAI's goal of a higher level of transparency, the user should be enabled to better understand and assess the capabilities and limitations of an AI in advance (ante-hoc) (Lepri et al., 2018; Liao et al., 2020). This transparency through XAI can be achieved in various ways, for example, based on text or visualization (Barredo Arrieta et al., 2020).

As people tend to be averse to machines' decisions (Dietvorst et al., 2015; Jussupow et al., 2020), and the opaque nature of AI can have a negative impact on trust (Hoffman et al., 2018), people might not trust candidate recommendations, rendering them ineffective for countering discrimination in hiring. Here, the emerging concept of XAI aiming to make AI use more transparent could be a promising method to increase trust in its recommendations (Thiebes et al., 2020).

Specifically, people tend to be cautious about technologies that are not interpretable or traceable (Barredo Arrieta et al., 2020), which could be reinforced by reports in the media stating that AI-based systems have led to discriminatory outcomes (Burke et al., 2021; Dastin, 2018). The goal of implementing XAI is to provide technical and contextual knowledge of how the underlying technology produces an output (Lepri et al., 2018; Mittelstadt et al., 2019), and XAI might also make it easier to identify and prevent unethical use of AI (Barredo Arrieta et al., 2020). We argue that XAI-induced transparency can increase reliance on the candidate recommendations of an AI-based system and result in users being more likely to follow the recommendations. However, as there are other studies indicating that providing transparency and domain knowledge can, in some cases, decrease trust of and reliance on a system (Dikmen & Burns, 2022; Hofeditz et al., 2021), it is difficult to determine if such an effect has a positive or a negative impact. Therefore, the following hypothesis is proposed:

H4: *Explaining an AI-based system moderates the effect of recommending candidates in an AI-based system on the selection of candidates.*

This hypothesis is divided into three sub-hypotheses based on the sensitive attributes used:

H4.1–H4.3: *Explaining an AI-based system moderates the effect of recommending foreign-race/older/female candidates in an AI-based system on the selection of foreign-race/older/female candidates.*

The research model is visualized in Fig. 1.

Research design

This study implemented a 2×2 between-subjects design and was conducted in the form of an online experiment due to health concerns during the COVID-19 pandemic. Quantitative and qualitative data were collected with a two-part online survey and with a task on a functional, interactive platform simulating an AI-based system for candidate management. In the task, participants were asked to select

Table 1 Experimental groups

		XAI-induced transparency (No I Yes)	
AI Recommendation	No	Group 1	Group 2
	Yes	Group 3	Group 4

Table 2 Questionnaires

Questionnaire	α	Author
Demographics		N/A
Big Five Inventory (BFI-10)	0.58–0.84	Rammstedt et al. (2013)
Affinity for Technology Interaction (ATI)	0.90	Franke et al. (2017)
Human Computer Trust Scale (HCTS)	0.83–0.88	Gulati et al. (2019)
NASA Task Load Index (NASA TLX)	0.83	Hart and Staveland (1988)
Ethics Position Questionnaire (EPQ)	0.80–0.83	Strack and Gennerich (2007)

suitable candidates for several job advertisements in a fictional organization.

Data collection took place between September 15, 2021, and October 20, 2021. The participants were equally distributed across four experimental groups (see Table 1). The four groups were varied in whether the participants received information about the functionality of the AI-based system ("XAI-induced transparency") and whether candidates with a sensitive attribute (regarding age, race, gender) were explicitly recommended by the AI-based system ("AI recommendation") on the candidate management platform. In more detail, we had AI recommendations as one varying factor and XAI-induced transparency as the other. In all groups, participants had to choose between two candidates each in 12 rounds per job and were (depending on the group) supported by AI recommendations, XAI, both, or neither. Among these 12 rounds, 6 represented the relevant rounds in which the candidates with sensitive attributes were recommended by the AI. The operationalization of our groups is explained in more detail in "Procedure."

Material

To investigate the impact of an AI-based system's recommendations on human decision-making in hiring, especially for typically disadvantaged candidates, a highly controllable and customizable environment was required. Previous literature has shown that users can evaluate AI-based systems if they believe that they are interacting with one (Hofeditz et al., 2021). This approach is related to the Wizard of Oz technique in which the functionality of the system is simulated by a human (the "wizard"). This technique can be used to test the interaction between humans and intelligent systems that cannot be easily implemented or realized with available resources (Weiss et al., 2009; Wilson & Rosenberg, 1988; Schoonderwoerd et al., 2022). Here, the system's functionality was not simulated by a human in real time but manually implemented prior to the experiment. Thus, this study did not develop a real AI-based system but a realistic, functional, and interactive prototype that simulated an AI-based system for candidate management. Specifically, we developed a candidate management platform called "nordflow" using the tool Bubble.io.[1] The presence of the AI-based system was simulated through a cover story and various user interface elements on the platform (e.g., loading screens indicating that the AI was analyzing applications). On the platform, participants navigated between three different job advertisements, reviewed applications for the respective position, and decided which candidates to invite. With this design, we followed the recommendations of Kuncel et al. (2014), who suggested using an algorithmic system based on a large number of datapoints to narrow a field of applicants before applying a human selection process for a few selected finalists. We placed emphasis on an intuitive user interface and realism of the platform to evoke realistic responses from the participants. The procedure section provides a more detailed overview of the platform and how the participants interacted with it. We tracked both quantitative data (participant decisions) and qualitative data (participant decision rationales). The former was used for hypotheses testing and answering our research questions, and the latter to gain richer insights into the participants' reasons for their decisions.

Furthermore, we used several questionnaires to assess different factors that might have influenced the results (Table 2). We used these controlling variables, as previous research has suggested considering a related combination in similar study contexts (Hofeditz et al., 2022a, 2022b; Mirbabaie et al., 2021a, 2021b, 2022).

The demographics questionnaire included questions on gender, age, employment status, educational attainment, and whether the participant had previous experience in HR. We then included the Affinity for Technology Interaction (ATI) scale to assess the tendency to actively engage in intensive technology interactions. The scale requires participants to rate their agreement with statements such as "I like testing the functions of new technical systems." We included a definition of "technical systems" to ensure a common understanding. Additionally, participants were asked to answer the Human Computer Trust Scale (HCTS), which was adapted to AI and the context of hiring; for example, "I think that Artificial Intelligence is competent and effective in selecting

[1] https://bubble.io/

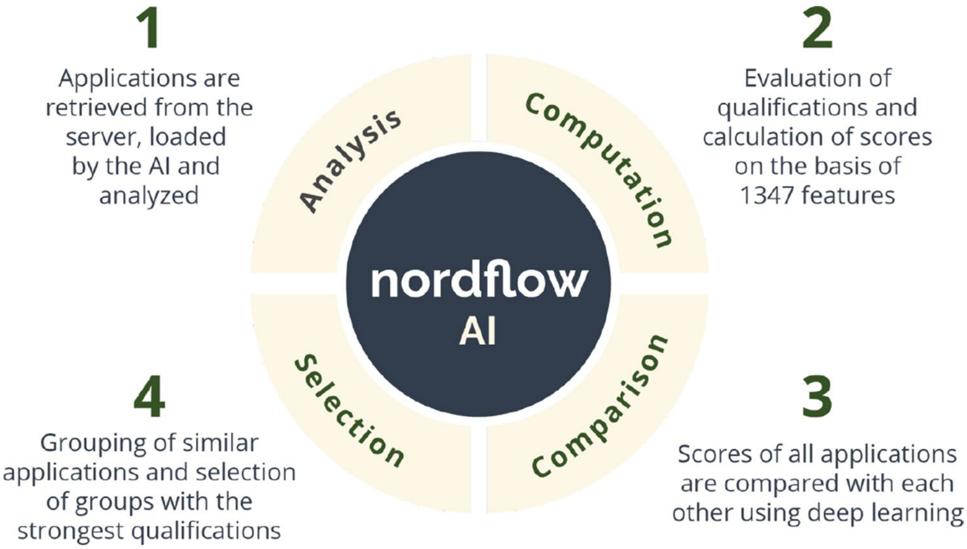
candidates." To ensure a common understanding of AI, we included a definition describing AI as "a system that can adapt independently to new situations and contents. It can solve problems and tasks that require a certain level of intelligence, as is typically present in humans." To measure the subjective cognitive load after interacting with the candidate management platform, we included the NASA Task Load Index (NASA TLX), consisting of questions such as "How mentally demanding was the task?" We excluded the scales for physical and temporal demand as those were not relevant for this study. The Ethics Position Question (EPQ) was used to measure ethical dispositions by asking for agreement to items such as "Risks to another should never be tolerated, irrespective of how small the risks might be." It was included last to avoid priming ethical behavior in the decision task. For all questionnaires, German translations or existing German versions were used, and all items were measured with a 7-point Likert scale. Following Oppenheimer et al. (2009), manipulation checks were implemented in the ATI and EPQ questionnaires to increase data quality and statistical power.

Procedure

First, participants received general information about the study and data protection and were asked to provide their written consent. It was specified that the study could only be completed on a desktop or laptop computer, and participants could not proceed if another device was used. Afterwards, participants were asked to answer the demographics, BFI-10, ATI, and HCTS questionnaires (Table 2) and were automatically assigned to one of four experimental groups (Table 1). Then, a cover story was presented to the participants stating that a (fictional) technology organization called "nordflow" had developed an AI-based system for candidate management and that the participants would be asked to interact with a prototype of that system. The participants were informed that the AI can pre-select a certain number of suitable candidates for different job advertisements by evaluating and rating their qualifications and fit for the job advertisement (visualized with star ratings). However, the AI cannot decide between applicants with particularly similar ratings. Therefore, the participants were asked to review sets of these similarly qualified candidates, decide whom to invite for an interview, and explain their decision. Thereby, participants were asked to consider the description of the job requirements (Fig. 7 in Appendix 5) and the qualification ratings of the candidates.

In the experimental groups with XAI-induced transparency (groups 2 and 4; Table 1), the participants additionally received an explanation of the functionality of the AI-based system. Specifically, the participants received a description in text form and a diagram showing the candidate selection and analysis process (see Fig. 2). In the text describing the AI-based system, the participants were informed that the AI-based system uses various algorithms in its calculations. It was emphasized that in the development of the AI, an important focus was placed on diversity and that the AI differentiates applicants on a variety of characteristics selected by a panel of experts (see Appendix 4 for details). The latter highlights that the foundation of data processing has also been verified and supported by external parties, which should lead to greater trust in the AI by the participants. It was emphasized that the AI's evaluation of candidates was based on objective criteria. Lastly, the participants were informed that the goal of the AI was to encourage decision makers to make more ethical decisions (i.e., decisions that enhance diversity) in candidate selection processes. Thus, participants in groups 2 and 4 received a high-level explanation of how the data is processed by the AI-based system,

Fig. 2 Process diagram of the candidate selection process (XAI-induced transparency)


1
Applications are retrieved from the server, loaded by the AI and analyzed
2
Evaluation of qualifications and calculation of scores on the basis of 1347 features
Analysis
Computation
nordflow AI
Selection
Comparison
4
Grouping of similar applications and selection of groups with the strongest qualifications
3
Scores of all applications are compared with each other using deep learning

which was intended to improve their understanding of why the AI selects and recommends certain candidates over others. This type of explanation relates to the first category of XAI approaches proposed in the taxonomy of Gilpin et al. (2018) and may increase user trust in the system's behavior. The participants in the other experimental groups did not receive this information.

Lastly, all participants were presented with a three-step tutorial explaining the platform's functionalities and instructions for using it. This included the job view (see Fig. 7 in Appendix 5), candidate selection view (Fig. 4), and screenshot of a text field for entering the decision rationale. The tutorial was adapted to the respective experimental group. After completing the tutorial, the participants were redirected to the candidate management platform.

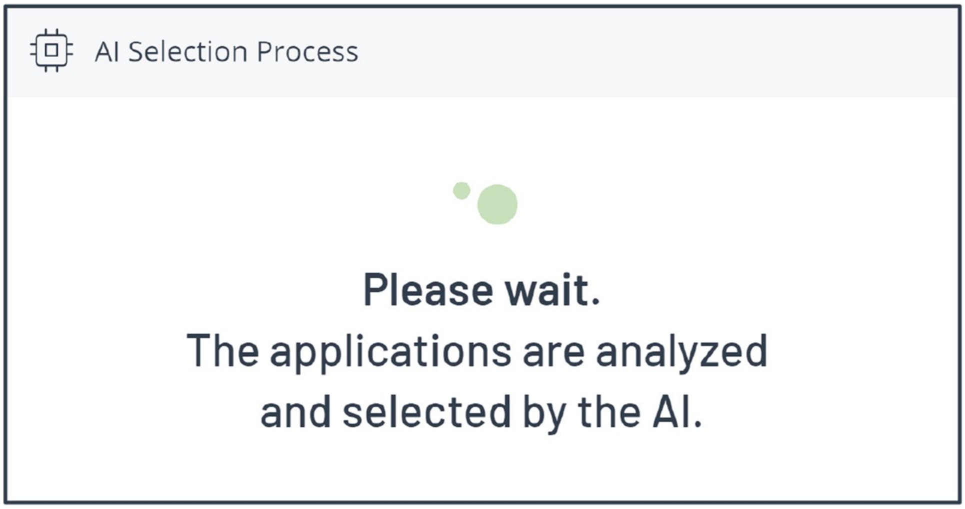
The "job view" of the platform showed four job advertisements, for three of which the participants should select candidates (Fig. 7 in Appendix 5). The job advertisements were identical for all participants but displayed in a randomized order. The participants were free to decide which job advertisement to start with. Each job advertisement was accompanied by a short description of the job. This description included references to different qualifications and was provided to ensure a common baseline for the participants' assessment of the candidate's qualifications.

After participants selected one of the three job advertisements by clicking on "Start Selection," an animated loading screen appeared that served to simulate the AI-based system's selection process (Fig. 3).

Next, the participants saw the "candidate selection view" (Fig. 4; round 5 for the job "IT administration"), which showed the personal attributes and qualification ratings for two candidates. It was ensured that the total qualification rating (sum of stars) was identical for the two candidates. The ratings for specific qualifications differed slightly between the candidates to enhance realism, to constantly test whether participants focused on the demographics, and to examine whether AI recommendations and XAI influenced this focus. The candidates' position, left or right, was randomized per participant to ensure that the recommendation was not always on the same side of the candidate window. Above the candidates, the description of the job advertisement was displayed as a reminder for the participants. The qualifications were visualized with a rating scale, as this makes different qualifications (e.g., different degrees) more comparable and reduces the influence of participants' individual preferences (e.g., for a specific language).

Participants were given the task of selecting candidates for the three job advertisements with sufficient applications. For each job advertisement, participants completed six rounds, comparing two candidates per round. Of the six rounds per job advertisement, three were "relevant rounds" that included one candidate with a sensitive attribute. In total, 36 candidates were created, of which 9 were of interest for the study (one candidate each in three relevant rounds per job). A complete list of candidates is included in Table 7 in the Appendix 2. The interplay between job advertisement, sensitive attribute, and relevant rounds is displayed in Table 3.

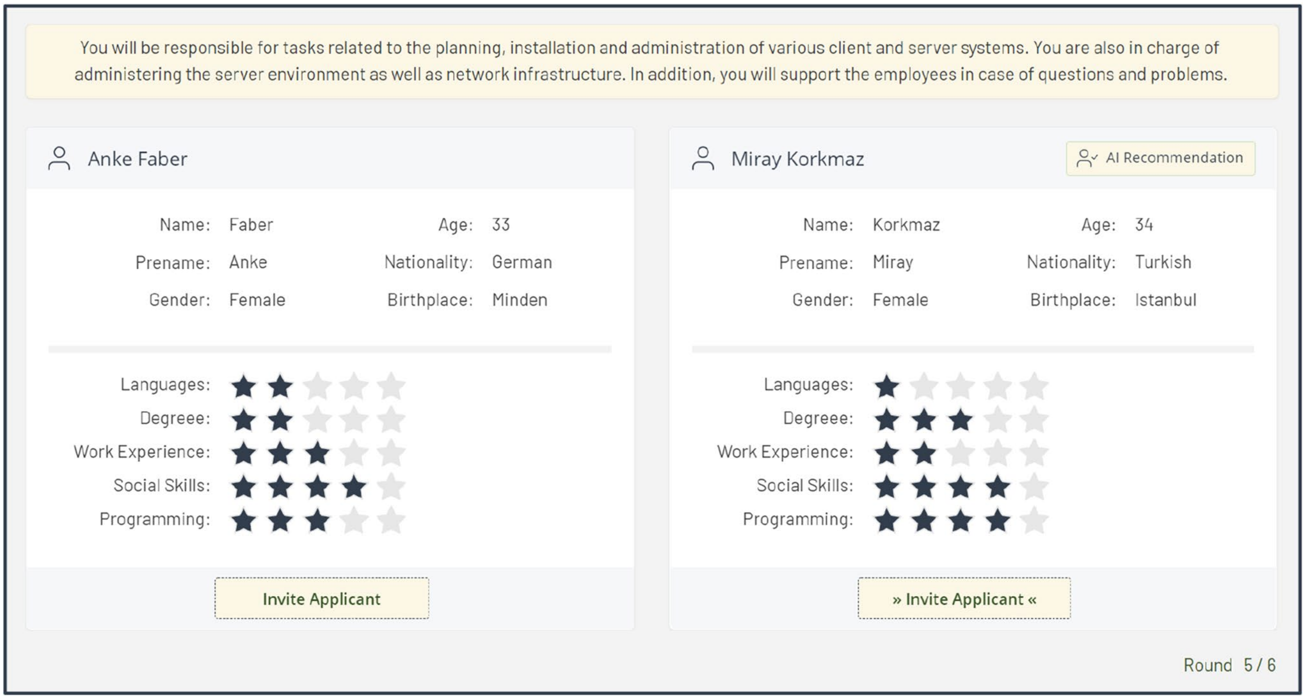
Only in experimental groups 3 and 4 were specific candidates recommended to the participant by the AI. Specifically, in the three relevant rounds per job, the candidate with a sensitive attribute was labeled with "AI Recommendation" in the upper right corner and the "invite applicant" button was complemented with small arrows (Fig. 4). In the nonrelevant rounds, the AI recommendation was "out of line," meaning that candidates without sensitive attributes might be recommended if they were more qualified. This approach was chosen to prevent the



AI Selection Process

Please wait.
The applications are analyzed
and selected by the AI.

Fig. 3 Loading screen simulating an AI-based system for candidate selection



You will be responsible for tasks related to the planning, installation and administration of various client and server systems. You are also in charge of administering the server environment as well as network infrastructure. In addition, you will support the employees in case of questions and problems.

Anke Faber

Name:	Faber	Age:	33
Prename:	Anke	Nationality:	German
Gender:	Female	Birthplace:	Minden

Languages:	★★☆☆☆
Degreee:	★★☆☆☆
Work Experience:	★★★☆☆
Social Skills:	★★★★☆
Programming:	★★★☆☆

Miray Korkmaz

Name:	Korkmaz	Age:	34
Prename:	Miray	Nationality:	Turkish
Gender:	Female	Birthplace:	Istanbul

Languages:	★☆☆☆☆
Degreee:	★★★☆☆
Work Experience:	★★☆☆☆
Social Skills:	★★★★☆
Programming:	★★★★☆

Fig. 4 Example of the "Candidate selection view" displaying job and candidate information, qualification ratings, and (in experimental groups 3 and 4) an AI recommendation

Table 3 Job advertisements round sequence

Job advertisement	Sensitive attribute	Round sequence					
		1	2	3	4	5	6
1 – IT-Administration	Race	R	R	D	D	R	D
2 – Project Management	Age	D	D	R	R	D	R
3 – Accounting	Gender	D	R	D	R	R	D

R relevant round with a candidate with a sensitive attribute, *D* nonrelevant round

participants from recognizing a pattern in the recommendations or candidates. The current round was displayed in the lower right corner to show the participants how far they had progressed.

To better understand why the participants selected a candidate, they were asked to enter the reason (as free text) for their decision after each choice. The cover story explained this to the participants by pointing out that supervisors wanted to track the reasons behind the decisions. Once the participants had completed all three job advertisements, they were directed back to the online survey. A complete overview of the order of the questionnaires, the content presented, and the information collected in this study is provided in Fig. 5.

Findings

Demographics

Individuals above the age of 18 years were eligible to participate in this study, and participants were recruited though SurveyCircle. Further restrictions for participation were not imposed. SurveyCircle is a research platform that helps European researchers recruit participants for online surveys and experiments. SurveyCircle's idea is to provide the opportunity to experience current online studies and actively support research in different disciplines through voluntary participation. As with SurveyCircle a completely representative sample cannot be guaranteed, we reached out

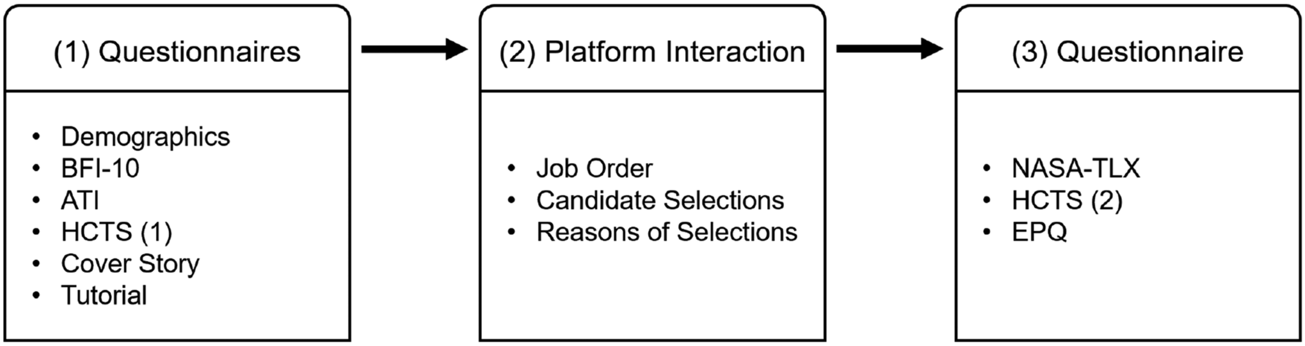


Fig. 5 Procedure

to additional participants via postings on LinkedIn, XING, and Facebook. It would be obvious to limit participation to people working in HR. However, we found that even among HR employees and within strategic HR management there are many differences between systems and HR philosophies (Lepak et al., 2004), which made it challenging to find a consistent group of HR employees while maintaining a large enough sample size. Also, employees in HR may already be aware that such a system could be used to test diversity in hiring, as this topic was already present in HR-relevant media, resulting in behavior unlike their natural decision behavior. Furthermore, AI-based systems augment people in the workplace in such a way that they can solve more complex tasks (Dellermann et al., 2019; Mirbabaie et al., 2021a, 2021b, 2021c). We therefore expect AI-based systems to enable increasingly more people in the future to perform tasks that were previously preserved by domain experts. Therefore, we decided to recruit not only current HR employees but also potential future leaders in companies as participants.

A total of 208 participants took part in the study. With 14 participants excluded for not providing reasons for their candidate selections on the interaction platform, 194 valid cases were included in the analysis. At the end of the survey, participants were asked whether they answered all the information honestly and in the best interests of the scenario (the participants were assured that their answer to this question would not put them at a disadvantage). No additional participants were excluded on the basis of this question. On average, participants spent 29 min completing the study. Participants ranged in age from 18 to 68 years ($M=28.28$, $SD=9.02$), of whom 126 were women (~65%) and 68 men. This approximates the real distribution of male and female employees working in HR in Germany, which is around 70% women and 30% men (Gorges, 2015). In addition, the sample shows that the participants were highly educated. A university degree was held by 68% of the participants, and a high school diploma or higher education entrance qualification by 23%. Furthermore, 67% of the participants reported being students, and 26% that they were employees. Between students ($M=25.05$, $SD=2.99$) and employees ($M=33.73$, $SD=11.06$), there was an age difference of almost 9 years. Moreover, among the employees, 68% reported having a university degree. Nearly one-third of the participants stated that they had experience in HR.

Quantitative findings

Effect of AI-based system's recommendations on candidate selection

To analyze whether the AI-based system's recommendations of typically disadvantaged individuals impact candidate selection in terms of race, age, and gender (H1–H3), unpaired *t*-tests were conducted. A candidate selection score, our dependent variable, was ratio scaled, and the independent variable was categorical with two groups. Furthermore, no outliers were identified. Except for normal distribution,

Table 4 Participants' candidate selections

		Job advertisement					
		Job 1		Job 2		Job 3	
No	Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	No recommendation and no XAI	2.11	0.759	1.91	0.905	1.45	0.829
2	No recommendation and XAI	1.81	0.970	1.89	0.759	1.47	0.830
3	Recommendation and no XAI	1.98	0.948	2.33	0.712	1.78	0.673
4	Recommendation and XAI	2.43	0.645	2.24	0.804	1.98	0.777

all requirements for the *t*-test were met. Table 4 shows the means and standard deviations of the four conditions and three job advertisements to provide an overview of the participants' candidate selections.

The relevant candidates with sensitive attributes (in terms of diversity) were coded 1, and the rest 0. The scores in Table 4 indicate (per condition and subdivided by job) the diversity in the participants' decisions (only for the rounds with candidates of minority groups). A 3 represents a decision toward selecting more diversity (relevant candidates with sensitive attributes were chosen), and a 0 represents selection of a candidate without sensitive attributes.

We then assigned a score for each participant per job. The values in Table 4 correspond to its mean across all participants, grouped by condition. Thus, in the example of Condition 1 regarding Job 1, participants selected an average of 2.11 relevant candidates with a sensitive attribute (higher age, female, or non-German). Comparing Condition 1 with 3 in Job 1, for example, the recommendations led to a reduction in more diverse candidate selection. Table 4 provides an overview of the candidates that allows a comparison of the conditions and jobs.

Regarding race (H1), participants who received the AI-based system's recommendations were less likely to select foreign-race candidates ($M = 1.98$, $SD = 0.948$) than those without recommendations ($M = 2.11$, $SD = 0.759$). There was no statistically significant difference between the candidate selection with recommendations and the group without recommendations, $t(96) = 0.722$, $p = 0.472$, $r = 0.075$. Thus, the first hypothesis was not supported. There was no significant effect of an AI-based system's recommendations on the selection of foreign-race candidates.

Regarding age (H2), participants who received recommendations from the AI-based system were more likely to select older candidates ($M = 2.33$, $SD = 0.712$) than those without recommendations ($M = 1.91$, $SD = 0.905$). There was a statistically significant difference between the candidate selections with recommendations and the group without recommendations, $t(96) = -2.555$, $p = 0.012$. The effect size is $r = 0.251$ and corresponds, according to Funder and Ozer (2019), to a medium-sized effect. The second hypothesis was supported, and there was a significant positive effect of an AI-based system's recommendations on the selection of older candidates.

Regarding gender (H3), participants who received the AI-based system's recommendations were more likely to select female candidates ($M = 1.78$, $SD = 0.673$) than those without recommendations ($M = 1.45$, $SD = 0.829$). The Levene test did not show homogeneity of variance ($p < 0.5$). Therefore, the Welch test was conducted. There was a statistically significant difference between the candidate selections with recommendations and the group without recommendations, $t(88.711) = -2.202$, $p = 0.03$. The effect size is $r = 0.214$ and corresponds, again according to Funder and Ozer (2019), to a medium-sized effect. The third hypothesis was supported, and there was a significant positive effect of the AI-based system's recommendations on the selection of female candidates.

In addition, moderating effects regarding gender, occupation, and HR experience were calculated using the PROCESS macro by Hayes (2018). The groups of students and employees were analyzed in terms of occupation, as they represented most of the participants. For the calculation of occupation, a new variable was calculated for each case, indicating the participant's respective group. The independence already mentioned for the *t*-test was also required for this procedure and was present, as it resulted from the experimental design. The relationship between the variables was not linear according to a visual inspection of the scatter plot after LOESS smoothing. However, the analysis continued, and a loss of statistical power was accepted. Bootstrapping was performed with 5,000 iterations and heteroscedasticity-consistent standard errors to calculate confidence intervals (CIs) (Davidson & MacKinnon, 1993). There is no centering of variables, as only the interaction effect is of interest.

Effect of XAI-induced transparency on candidate selection

To analyze whether the interaction between XAI and the AI-based system's recommendations significantly predicted participants' candidate selections, moderation analyses using the PROCESS macro by Hayes (2018) were conducted. Bootstrapping was performed with 5,000 iterations and heteroscedasticity-consistent standard errors to calculate CIs (Davidson & MacKinnon, 1993). The relationship between the variables was not linear for any of the XAI hypotheses according to a visual inspection of the scatter plot after LOESS smoothing. However, the analysis continued, and a loss of statistical power was accepted. There is no centering of variables as only the interaction effect, the influence of XAI, is of interest. To perform the moderation analysis, two new variables were calculated from the stimulus variable that represents all four groups. Two variables were created indicating whether participants received a condition including recommendation (regardless of XAI; $n = 100$) or XAI (regardless of recommendation; $n = 96$).

The overall model regarding the selection of foreign-race candidates was significant $F(3, 190) = 5.46$, $p = 0.001$, predicting 6.88% of the variance. The moderation analysis showed that XAI significantly moderated the effect between the AI-based system's recommendation and the selection of foreign-race candidates: $\Delta R^2 = 4.66\%$, $F(1, 190) = 9.33$, $p = 0.002$, 95% CI[0.279, 1.236]. Thus, Hypothesis 4.1 was confirmed.

The overall model regarding the selection of older candidates was significant $F(3, 190) = 4.04$, $p = 0.008$,

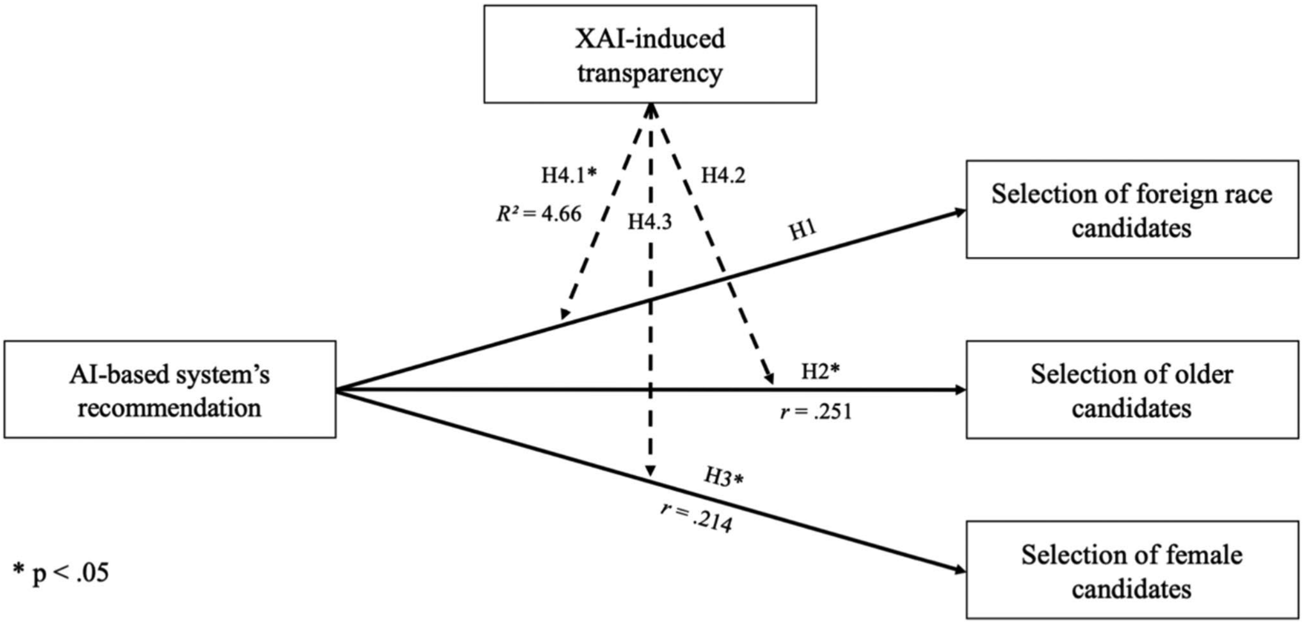

XAI-induced transparency
H4.1*
R² = 4.66
H4.3
H4.2
H1
AI-based system's recommendation
Selection of foreign race candidates
H2*
r = .251
Selection of older candidates
H3*
r = .214
Selection of female candidates
* p < .05

Fig. 6 Summarized findings

predicting 5.8% of the variance. However, the moderation analysis did not show that XAI significantly moderated the effect between the AI-based system's recommendation and the selection of older candidates: $\Delta R^2 < 0.01\%$, $F(1, 190) = 0.084$, $p = 0.772$, 95% CI[-0.518, 0.384]. Hypothesis 4.2 was not confirmed.

The overall model regarding the selection of female candidates was significant $F(3, 190) = 4.96$, $p = 0.002$, predicting 7.72% of the variance. The moderation analysis did not show that XAI significantly moderated the effect between the AI-enabled candidate recommendation system and the selection of female candidates: $\Delta R^2 < 0.01\%$, $F(1, 190) = 0.587$, $p = 0.444$, 95% CI[-0.276, 0.613]. Hypothesis 4.3 was not confirmed.

As only one sub-hypothesis showed significance, Hypothesis 4, which states that XAI moderates the effect between an AI-based system's recommendations and candidate selections, could not be confirmed.

To summarize the findings, Fig. 6 shows the quantitative results for all hypotheses. Further results, such as an overview of the BFI-10, can be found in the Appendix.

Qualitative findings: Reasons for participants' selection behavior

The participants were asked the following question after each selection: "Why did you select this candidate?" Their reasons for selecting candidates on the platform is evaluated in the following sections. The dataset consists of 1,746 fields (194 participants × 3 jobs × 3 reasons), including keywords, sentences, and short argumentations. To gain insights into the reasons for selections, we conducted qualitative content analysis according to Mayring (1994).

Content analysis allowed us to summarize and reduce the participants' reasons to their essential message. The coding categories were derived inductively and included the five different qualifications of the candidates (i.e., languages, degree, work experience, social skills, and programming/methods/software skills) and a general qualification category for those cases in which the participants did not specify which qualification contributed to the decision. Furthermore, the categories included the three sensitive attributes (i.e., race, age, gender), each with the sub-categories of "sensitive attribute preferred," "non-sensitive attribute preferred," or "unspecified." Additionally, we included a category for coding whether the participants mentioned that they followed the AI recommendation. Lastly, we included the categories "subjective" for personal reasons, "ethical" for general comments about diversity without a specific reference to race, age, or gender, and "excluded" for comments that did not specify any reasons. The coding was conducted by two of the authors, who discussed the coding at several points in the process to decide on ambiguous cases. Comments could be assigned to multiple categories. Table 5 shows the derived categories with their relative occurrence in percentages.

Regarding the percentages for the sensitive attributes of race, age, and gender, all mentions of these attributes in the comments were considered. Thus, the percentage displayed includes all comments referring to a sensitive attribute, regardless of whether these are positively or negatively framed. This was done because considering these sensitive attributes in a hiring process could already be seen as

Table 5 Selection reasons mentioned by participants with and without HR experience in percentages

	L	*D*	*WoE*	*SoS*	*P/M/S*	*Q*	*Rec*	*Race*	*Age*	*Gender*
Job 1	7.56	10.31	32.30	16.84	50.34	16.15	13.00	1.37	1.20	1.89
HR	9.70	8.48	32.73	18.79	49.70	14.55	9.80	1.82	0.00	2.42
No HR	6.71	11.03	32.12	16.07	50.60	16.79	14.65	1.20	1.68	1.68
Job 2	2.06	6.19	31.44	29.38	31.44	24.40	13.33	0.00	12.89	0.00
HR	0.61	4.85	30.30	26.06	23.03	30.91	8.82	0.00	18.18	0.00
No HR	2.64	6.71	31.89	30.70	34.77	21.82	15.66	0.00	11.03	0.00
Job 3	4.81	22.34	20.62	18.73	31.27	23.02	17.00	0.00	1.55	4.64
HR	1.82	22.42	21.21	21.21	28.48	23.03	15.69	0.00	0.00	3.03
No HR	6.00	22.30	20.38	17.75	32.37	23.02	17.68	0.00	2.16	5.28

55 participants had HR experience, 139 had no HR experience

L language, *D* degree, *WoE* work experience, *SoS* social skills, *P/M/S* programming / methods / software skills, *Q* qualification, *Rec* recommendation; *Job 1* IT Administration, *Job 2* Project Management, *Job 3* Accounting

a form of discrimination. For an explorative comparison, Table 5 distinguishes the reasons provided by participants with domain knowledge in HR compared to those without domain knowledge in HR.

For the first job, "IT-Administration," approximately half the participants mentioned "programming" as a reason for selecting candidates. For the other jobs, the reasons for selection were more evenly distributed across different skills. Furthermore, while race and gender were rarely explicitly mentioned in the comments, relatively more comments addressed the age of the participants. Interestingly, more participants with HR experience mentioned age than those without HR experience. Lastly, it is interesting that participants with HR experience mentioned the AI recommendation less frequently as a reason for their selection compared to participants without HR experience.

We also examined the comments on race, age, and gender in more detail. All but one comment coded in the gender category expressed a preference for female candidates, mentioning, for example, the quota for women (ID 1771; ID 1565), that in the case of similar qualifications, women should be preferred (ID 1739), that women should be supported in certain disciplines, such as IT (ID 1848), or that a "woman is always good for team morale" (ID 1492).

With regard to race, several participants expressed a preference for non-Turkish candidates, for example, "I would not invite a Turk" (ID 1405) or "Turkish, but still more IT experience" (ID 1492) or stated "German" as the only justification for their candidate selection (ID 1385). However, one participant emphasized positive aspects of increasing diversity with new hires, stating "intercultural, therefore access to other resources" (ID 1749).

Lastly, almost all participants who commented on a candidate's age preferred younger candidates, stating, for example, that another candidate has "more time before retiring" (ID 1395). Several participants connected age to the ability to learn quickly, which compensated for methodological skills that they were currently lacking. For example, "methodological competence can possibly be further developed due to his age" (ID 1692). The only pro-older candidate comment was "Older women should be supported" (ID 1848)..

Discussion

Why AI recommendations might not reduce race-based discrimination in hiring

The participants' selection of foreign-race candidates was not in line with existing literature on racial discrimination in hiring, which has found that race continues to be one of the main sources of discrimination in hiring today and that Turkish individuals are especially discriminated against in Germany (Baert, 2018; Quillian et al., 2017, 2019; Zschirnt & Ruedin, 2016). This was generally not the case in our study, as the control group without recommendations and explainability showed a high number of selections of foreign-race candidates. However, if the AI-based system recommended a foreign-race candidate, the candidate was less likely to be selected. Thus, the AI-based system's recommendations did not increase the selection of foreign-race candidates, and Hypothesis 1 was not supported. Instead, participants who received the recommendations tended to select fewer foreign-race candidates than participants who did not receive recommendations. As the BFI-scores of the participants in the experimental groups receiving AI recommendations did not differ significantly from the groups without AI recommendations (see Table 8 in Appendix 5), this difference does not appear to result from personality differences of the participants in the respective groups.

We suggest that the implemented AI recommendation did not work as assumed because of algorithmic aversion (Berger et al., 2021; Dietvorst et al., 2015; Ochmann et al., 2021). Although algorithmic aversion usually occurs if people with domain knowledge can select between a human and an algorithm recommendation (Dietvorst et al., 2015), some previous research has suggested that there are cases of algorithm aversion occurring even if there is no human recommender alternative (Bigman et al., 2021). The qualitative analysis also provides evidence for this, suggesting that participants with domain knowledge in HR relied less on the AI recommendations compared to participants without domain knowledge. Accordingly, there is a possibility that the participants were indeed averse to AI, which led to a rejection of the recommended candidate and an increased selection of German candidates. This could be another indicator that algorithmic aversion can occur even without offering a human alternative, as suggested by Bigman et al. (2021). To avoid possible aversion, AI-based systems might be used to some extent as part of new task designs that balance human and system characteristics through mutual delegation (Baird & Maruping, 2021). Another approach to mitigating aversion is to get affected individuals better involved in the AI adaptation process as part of organizational learning (Wijnhoven, 2021). On the other hand, the lower reliance on the AI recommendations of some of the participants might also be due to a higher degree of self-confidence in selecting the right candidate, as recent research has found that self-confidence influences the adoption or rejection of AI augmentation (Chong et al., 2022).

As the reasons provided by participants with and without HR practice differed only slightly and participants could not select between a human and an algorithmic advisor, further explanations than aversion and self-confidence need to be taken into account. Considering the qualitative results, the low number of reasons given based on a candidate's race suggests that participants did not pay much attention to the race of candidates or that they were trying to be as objective as possible in decision-making. This could indicate that the selection decisions were, in fact, predominantly made based on qualifications. While the overall qualification (sum of stars) was identical for the candidates, the individual scores for qualifications differed slightly. However, it is also possible that participants were not aware of or avoided mentioning the role of the candidate's race in their selection, either because they were not aware of their own biases or because they did not want to admit them (i.e., the answers might be subject to a social desirability bias).

Examining the participants' reasons in more detail, we found that programming experience was the most frequently mentioned reason for the decision in the first job round in which race was the sensitive attribute. We therefore assume that the reason for not finding an effect of AI recommendation on selecting foreign-race candidates could be that the majority of our participants were sure that programming skills comprised the most important criterion for the role of IT administrator even though other skills were mentioned in the job description. Thus, an AI recommendation does not seem to be effective when there is already a clear qualification-based indicator for a decision. This could be explained by a high level of confidence that results in the avoidance of following AI recommendations. The result of the personality test (BFI-10) could also be used to explain the lack of evidence for discrimination in the initial candidate selection. A high level of openness (see Table 8 in Appendix 5) indicates that the participants think unconventionally and are open to new things (John & Srivastava, 1999). This could foster consideration of the overall qualification of the candidates regardless of their demographics and thus, result in less discrimination. However, we did not find systematic differences in personality between the relevant groups. When considering the results for Hypotheses 2 and 3, it becomes apparent that the AI-based system's recommendations can also work as expected in cases where participants perceive less clear qualification-based criteria for job profiles than was the case with programming for the first job.

AI recommendations can reduce age- and gender-based discrimination

When examining candidate selection in the control group, it becomes apparent that compared to the attributes of race and age, the participants selected female candidates considerably less often. This reinforces the evidence in the literature regarding discrimination against female candidates in hiring (Baert, 2018; Carlsson & Sinclair, 2018; Kübler et al., 2018). The qualitative data rather signaled that if participants mentioned gender as a reason for selection, they emphasized positive (yet partially stereotypical) aspects of hiring women (e.g., being good for team morale). This suggests that the negative discrimination against women shown by the quantitative results happened unconsciously or that participants deliberately concealed the discrimination.

The quantitative finding that AI recommendations can increase the selection of older and female candidates (H2 and H3) can be further strengthened by the qualitative results, which reveal that 13% and 17%, respectively, of the participants who received recommendations mentioned it as a reason for their candidate selection. In addition, participants stated that they used the recommendations to make decisions in cases of uncertainty. Regarding the second hypothesis, where we considered whether an AI-based system's recommendations impact the selection of older candidates, it was supported and showed a medium-sized effect. This implies that the recommendations led to more frequent selection of older candidates. These findings are strengthened by the qualitative findings, in which the participants mentioned the recommendation as a reason for their selection. Furthermore, the participants' ethical position (EPQ) indicated that they possessed rational and diverse candidate selection behavior. The findings for the sensitive attribute age (H2) are also in line with current literature regarding discrimination, which shows that older candidates are subject to discrimination in hiring (Baert, 2018; Lössbroek et al., 2021; Neumark et al., 2017; Zaniboni et al., 2019). In summary, the AI-based system's recommendations positively influenced the participants' selection decisions for older and female candidates.

Comparing the participants with and without HR experience, those with HR experience mentioned age more often as a reason for their decision. If we assume that a candidate was selected mainly because of their age and not because of a certain skillset, we consider this a case of age discrimination (Baert, 2018; Neumark, 2021; Richardson et al., 2013; Zaniboni et al., 2019). However, AI-based recommendations showed a positive effect on the selection of an older candidate for both groups.

The role of XAI in AI-based systems for hiring

Our findings suggest that XAI-induced transparency, that is, providing participants information about the functionality of the AI-based system, did not moderate the effect of the system's recommendations on the selection of older and female candidates (H4.2 and H4.3 rejected). It appears that emulating the processing of an AI-based system by providing a high-level explanation of the input–output relation of the data did not – as would be expected based on the suggestions of Gilpin et al. (2018) – increase the participants' trust in and acceptance of the system's recommendations. Thus, these findings seem to challenge expectations highlighting the effectiveness of and general need for XAI (Adadi & Berrada, 2018; Dwivedi et al., 2021; Gunning et al., 2019). One reason for this might be that there is a wide range of XAI types (see, e.g., Giudotti et al. (2018) for an overview) and that a different XAI type would have been more suitable to support the target audience. However, as XAI-induced transparency positively moderated the selection of foreign-race candidates (H4.1 supported), the effectiveness of XAI might also depend on the content of the decision task. Previous research has already emphasized that a successful application of XAI depends on various quality criteria, such as fidelity, generalizability, explanatory power, interpretability, comprehensibility, plausibility, effort, privacy, and fairness, depending on the target group (Meske et al., 2022). Here, with H4.2 and H4.3 not being supported and H4.1 being supported, our findings suggest that not only quality criteria and XAI type but also the content of the decision task need to be considered.

With these findings, we addressed the research gap identified by Adadi and Berrada (2018), who argued that the role of humans in XAI is inconclusive and can only be attributed to undiscovered influencing factors. We provided empirical evidence for the context of discrimination in hiring and tested XAI in the context of participants' ethical position and personality traits. In addition, our findings suggest that the content for achieving XAI-induced transparency should be individually adaptable to user qualifications. This is in line with Shin (2021), who argued that algorithmic experience in AI needs to be addressed in practice and that heuristics and cognitive processes need to be incorporated into the design of these algorithms, making them user-centric. Furthermore, based on our findings, more research is needed regarding the mechanisms of XAI on humans and their influencing factors, which was also one of the research opportunities outlined by Meske et al. (2022) for XAI in information systems. In addition, we provided empirical evidence on how a higher degree of transparency leads to better understanding of potentially undesired practices in the offline world (e.g., gender bias and discrimination), which was mentioned as a promising research direction by Meske et al. (2022). We addressed both knowledge on XAI in the context of individual attributes and knowledge on how XAI and transparency can lead to less discrimination and bias in hiring.

Limitations and further research

The study adopted a fairly broad, high-level type of XAI in which participants received a general explanation about the processing of data in the system as well as its goal of augmenting decision-making in hiring to reduce discrimination. However, there are many other, more technically detailed XAI approaches that could prove (more) effective in this context (see, e.g., Adadi & Berrada, 2018 or Gilpin et al., 2018). While this study focused on a target audience (and a corresponding sample) of non-AI experts, we acknowledge that this might be an insufficiently detailed characterization of HR professionals. In addition, a relatively large number of participants were educated, female, and from Germany, and only about one-third of the participants reported prior HR experience. Therefore, the findings are subject to limited generalizability to HR professionals.

The candidate management platform was designed to resemble prevalent AI-based systems for this purpose; however, the findings might not be generalizable to other platforms in this domain. The overall qualification of the candidates (sum of stars) was identical for both candidates in each round; the star ratings for specific qualifications differed between the two candidates. While this was necessary to gain insights on participants' tendency to consider demographic information for deciding between candidates, it introduced the risk that the participants' perceived relevance of certain qualifications for a job influenced their selection. Also, the effects might differ if participants had to select candidates from a larger pool of candidates on the platform rather than making a choice based on a direct comparison of two candidates. Lastly, the cultural context in which the platform is deployed might make a difference as for example the influence of AI recommendations could be more pronounced in highly technology-affine societies.

Addressing these limitations, future research could explore the (dis-)advantages of different types of XAI from the perspective of HR professionals in greater depth. It would be interesting to conduct the study in a real HR environment and limit participation to experienced HR employees. As the sensitive attributes leading to discrimination might differ depending on contextual factors (e.g., culture) or individual factors (e.g., characteristics of the decision maker), future studies should aim to explore the effects of AI recommendations and XAI with different sensitive attributes (e.g., disability) and a diverse group of HR professionals. Furthermore, to dive deeper into possible causes for the observed candidate selection behavior and the effectiveness of AI recommendations and XAI, future research could measure algorithmic aversion, automation bias, cognitive load, or the effect of mistrust disposition. Future research should consider directly measuring these aspects in the context of XAI and AI recommendations for candidate selection and examine possibilities to mitigate aversion, for example, by incorporating AI-based information systems as part of new task designs that balance human and systemic characteristics through mutual delegation and through organizational learning processes with strong stakeholder participation in AI adoption. Additionally, to improve generalizability, future research could investigate XAI and AI recommendations on different types of candidate management platforms and in alternative deployment contexts (e.g., other countries).

Lastly, we focused on the point of view of the recruiter and not on those who are affected by discrimination or bias in hiring. Future research needs to go a step further and, for example, follow a discourse ethics approach based on that of Mingers and Walsham (2010) by also involving other stakeholders in the debate about diversity in XAI-based recommendations.

Contribution to research

The findings of this study contribute to research on augmenting human decision-making with AI-based systems in several ways. First, we showed that in decision-making scenarios with no clearly preferable option, providing AI recommendations and XAI can influence decision-making and potentially reduce discrimination in hiring. Second, our findings suggest that a clear association between a qualification-based criterion and a decision outcome limits the impact of AI recommendations on decision-making. Third, our exploratory analysis indicated that participants with domain knowledge did not behave differently in response to AI recommendations and/or XAI than participants without domain knowledge. Fourth, we open a new field of research regarding the combination of XAI and AI-based system recommendations to augment decision-making in the context of hiring.

We also contribute to the literature on XAI by empirically testing the influence of XAI on the effectiveness of augmenting decision making with an AI-based system in the context of hiring. As the effects of XAI differed for the sensitive attributes, our findings suggest that, in addition to quality criteria and target groups (Meske et al., 2022), the content or context of the decision plays a role in the impact of XAI.

Furthermore, this research extends the literature concerning the reduction of discrimination in hiring (e.g., Foley & Williamson, 2018; Krause et al., 2012) and presents recommendations regarding an AI-based system as a promising approach for reducing discrimination against older and female candidates in hiring. Moreover, the findings argue for the positive benefits of using AI to reduce discrimination and bias, complementing the literature that discusses the ethical issues of AI in hiring (Lepri et al., 2018; Raghavan et al., 2020). Finally, the study contributes to broadening the understanding of AI in society by demonstrating a new beneficial use case of applying XAI to reduce discrimination in hiring.

Contribution to practice

This research also provides practical implications for stakeholder groups working with XAI, such as AI managers, AI developers, AI users, and individuals affected by AI-based system decisions and recommendations.

On the one hand, the study contributes to increasing general welfare by examining an important topic for society and electronic markets. Thus, our findings might lead to greater diversity in future workforces and positively affect individuals with sensitive attributes who are subject to AI-based system recommendations. For example, recruiters as AI users can augment their decision making with similar systems, reflect on their (potential) biases, and better understand the reasons for AI-based system recommendations through XAI. On the other hand, this research can draw the attention of organizations and AI managers to the issue of discrimination remaining an important problem in hiring. Furthermore, the platform conceptualized and developed in this research can be a starting point for developing a system for training HR staff on discrimination in hiring. XAI and recommendations from AI-based systems can be effective, but they may require further action from the organization to achieve diverse hiring in the long term.

For practical application purposes, XAI might be successful in areas where users are in more frequent contact with the technology. Complementing XAI, the implementation of AI recommendations might be a suitable method to realize the AI's purpose of, in this case, countering discrimination in hiring. Moreover, in general, educating different stakeholders of XAI about AI's potential dangers and benefits would be advisable to reduce prejudice and fear and increase general acceptance of AI.

From an organizational perspective, a question arises as to the overall benefits of XAI for their business. Not every organization that uses an AI-based system necessarily needs to understand the reasons for its outcomes. In addition, some algorithms must be developed from scratch to allow for the ability to explain the processes and reasons for decisions afterward. This leads to an immense amount of work, which may not justify the perceived benefits of XAI in every context. Therefore, incentives are needed that could counteract some of the barriers to the implementation of XAI to ensure more diversity and fewer biases through XAI and AI based-system decisions and recommendations.

Conclusion

In summary, our findings suggest how recommendations by an AI-based system for hiring, combined with an XAI approach, can be applied on a candidate management platform to achieve greater transparency and diversity. It appears that AI recommendations are sufficient to cause participants to reconsider their decision-making or to draw attention to sensitive attributes. While our findings might not generalize to other AI-based systems or candidate management platforms, we found that AI recommendations encouraged decision makers to select more female and older candidates. However, the recommendations also resulted in fewer selections of foreign-race candidates, which might be due to algorithmic aversion caused by overly obvious recommendations based on sensitive attributes. Furthermore, while explainability moderated the effect of AI recommendations on the selection of foreign-race candidates, our findings cannot unreservedly support the positive impact of explainability on the effect of AI recommendations on selection behavior. However, our findings overall suggest that AI recommendations can reduce discrimination in hiring decisions. We further conclude that XAI helped reduce reactance and aversion caused by recommendations that were too obviously perceived as an influencing factor by our participants. The XAI appeared to have different effects for the same target group and the same quality criteria, highlighting the importance of considering the content of the decision task.

Appendix 1

Table 6 Reasons for discrimination in hiring

Discrimination type	Disadvantaged groups	Exemplary sources
Race	Minority races and national origins	(Baert, 2018; Weichselbaumer, 2016)
Gender	Females or mothers	(Baert, 2018; Ruffle & Shtudiner, 2015)
Age	Too young or old	(Baert, 2018; Neumark, 2021; Richardson et al., 2013; Zaniboni et al., 2019)
Religion	Minority religions	(Baert, 2018; Weichselbaumer, 2016)
Disability	Any disability	(Baert, 2018; Stone & Wright, 2013)
Sexual orientation	Non-heterosexual orientation	(Baert, 2018; Weichselbaumer, 2016)
Physical appearance	Low attractiveness, females with high attractiveness	(Baert, 2018; Stone & Wright, 2013)

Appendix 2

Table 7 Candidate profiles

Job	Round	Type	Prename	Name	Gender	Age	Birthplace	Nationality	Lang	Deg	Exp	Social	Extra
1	1	0	Dirk	Bauer	Male	41	Solingen	German	3	4	3	3	2
1	1	1	Mustafa	Özdemir	Male	43	Tarsus	Turkish	2	3	4	3	3
1	2	1	Ecrin	Şahin	Female	37	Adana	Turkish	4	3	1	2	4
1	2	0	Stephanie	Jung	Female	39	Kerpen	German	3	2	2	4	3
1	3	99	Paulina	Krasny	Female	37	Breslau	Polish	1	3	2	2	4
1	3	98	Wolfgang	Fischer	Male	54	Wittlich	German	3	2	5	1	1
1	4	99	Mathias	Eisenberg	Male	42	Ulm	German	1	2	3	1	3
1	4	98	Dennis	Zimmer	Male	37	Rosenheim	German	3	4	2	3	2
1	5	1	Miray	Korkmaz	Female	34	Istanbul	Turkish	1	3	2	4	4
1	5	0	Anke	Faber	Female	33	Minden	German	2	2	3	4	3
1	6	99	Maria	van Dijk	Female	39	Venlo	Dutch	4	2	4	3	3
1	6	98	Jürgen	Fuhrmann	Male	51	Bramsche	German	2	3	3	5	2
2	1	99	Luisa	Engel	Female	56	Marburg	German	3	4	4	2	2
2	1	98	Sophia	Thalberg	Female	47	Lörrach	German	2	3	3	4	5
2	2	98	Oskar	Borkowski	Male	37	Lissa	Polish	2	3	3	1	5
2	2	99	Simone	Wulf	Female	33	Leipzig	German	5	5	1	4	3
2	3	0	Sven	Kuster	Male	33	Flensburg	German	2	4	3	2	4
2	3	1	Thomas	Ackermann	Male	54	Tübingen	German	2	2	4	4	3
2	4	1	Monika	Zimmer	Female	57	Niebüll	German	3	4	4	2	3
2	4	0	Lisa	Schaefer	Female	36	Babelsberg	German	2	5	3	4	2
2	5	98	Philipp	Neumann	Male	54	Essen	German	2	3	4	5	4
2	5	99	Katja	Weiß	Female	36	Halle (Saale)	German	4	2	3	5	2
2	6	0	Martin	Bach	Male	39	Nürnberg	German	5	3	2	4	1
2	6	1	Patrick	Lehmann	Male	51	Hilden	German	3	3	3	2	4
3	1	99	Robin	Winkler	Male	51	Wesel	German	3	5	3	2	5
3	1	98	Jannik	Grunewald	Male	48	Bocholt	German	3	2	4	3	4
3	2	0	Christian	Nacht	Male	37	Bayreuth	German	2	3	3	5	2
3	2	1	Katharina	Decker	Female	36	Celle	German	3	4	2	3	3
3	3	99	Laura	Fischer	Female	34	Paderborn	German	2	1	2	3	2
3	3	98	Aylin	Öztürk	Female	45	Mersin	Turkish	4	2	4	2	3
3	4	0	Arne	Meyer	Male	44	Halberstadt	German	4	4	4	2	3
3	4	1	Karin	Richter	Female	42	Fulda	German	3	5	4	3	2
3	5	1	Sophia	Ostermann	Female	34	Dresden	German	2	3	2	4	2
3	5	0	Dominik	Braun	Male	33	Rathenow	German	3	4	1	2	3
3	6	98	Anna	Iwanow	Female	57	Krasnojarsk	Russian	3	4	5	3	2
3	6	99	Aaron	Becker	Male	39	Koblenz	German	2	3	3	4	5

Column Type: 0 = relevant; 1 = relevant with recommendation; 98 = non-relevant with recommendation; 99 = non-relevant. Lang. = Languages; Deg. = Degree; Exp. = Work Experience; Social = Social Skills; Extra = Dynamic Field

Appendix 3

Questionnaires

All questionnaires are based on a 7-point Likert scale.

Big Five Inventory (BFI-10)

Bitte geben Sie den Grad Ihrer Zustimmung zu folgenden Aussagen an.

1. Ich bin eher zurückhaltend, reserviert.
2. Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.
3. Ich bin bequem, neige zur Faulheit.
4. Ich bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen.
5. Ich habe nur wenig künstlerisches Interesse.
6. Ich gehe aus mir heraus, bin gesellig.
7. Ich neige dazu, andere zu kritisieren.
8. Ich erledige Aufgaben gründlich.
9. Ich werde leicht nervös und unsicher.
10. Ich habe eine aktive Vorstellungskraft, bin fantasievoll.

Affinity for Technology Interaction (ATI)

Bitte geben Sie den Grad Ihrer Zustimmung zu folgenden Aussagen an.

Mit „technischen Systemen " sind sowohl Apps und andere Software-Anwendungen als auch komplette digitale Geräte (z.B. Handy, Computer, Fernseher, Auto-Navigation) gemeint.

1. Ich beschäftige mich gern genauer mit technischen Systemen.
2. Ich probiere gern die Funktionen neuer technischer Systeme aus.
3. In erster Linie beschäftige ich mich mit technischen Systemen, weil ich muss.
4. Wenn ich ein neues technisches System vor mir habe, probiere ich es intensiv aus.
5. Ich verbringe sehr gern Zeit mit dem Kennenlernen eines neuen technischen Systems.
6. Es genügt mir, dass ein technisches System funktioniert, mir ist es egal, wie oder warum.
7. Ich versuche zu verstehen, wie ein technisches System genau funktioniert.
8. Es genügt mir, die Grundfunktionen eines technischen Systems zu kennen.
9. Ich versuche, die Möglichkeiten eines technischen Systems vollständig auszunutzen.

Human Computer Trust Scale (HCTS)

Bitte geben Sie den Grad Ihrer Zustimmung zu folgenden Aussagen an.

Eine künstliche Intelligenz (KI) lässt sich als System beschreiben, dass die Fähigkeit besitzt, sich selbstständig an neue Situationen und Inhalte anzupassen. Es kann Probleme lösen und Aufgaben erledigen, die ein gewisses Maß an Intelligenz erfordern, wie sie typischerweise bei Menschen vorhanden ist.

1. Ich glaube, dass der Einsatz einer künstlichen Intelligenz negative Folgen haben könnte.
2. Ich glaube, ich muss vorsichtig sein, wenn ich eine künstliche Intelligenz verwende.
3. Es ist riskant, mit einer künstlichen Intelligenz zu interagieren.
4. Ich glaube, dass eine künstliche Intelligenz in meinem besten Interesse handeln wird.
5. Ich glaube, dass eine künstliche Intelligenz ihr Bestes tun wird, um mir zu helfen, wenn ich Hilfe benötige.
6. Ich glaube, dass eine künstliche Intelligenz daran interessiert ist, meine Bedürfnisse und Vorlieben zu verstehen.
7. Ich denke, dass eine künstliche Intelligenz bei der Auswahl von Bewerber:innen kompetent und effektiv ist.
8. Ich denke, dass eine künstliche Intelligenz ihre Rolle als Instrument zur Bewerberauswahl sehr gut erfüllt.
9. Ich glaube, dass eine künstliche Intelligenz über alle Funktionen verfügt, die ich von einem Hilfsmittel zur Bewerberauswahl erwarten würde.
10. Wenn ich eine künstliche Intelligenz verwende, denke ich, dass ich mich vollständig auf sie verlassen kann.
11. Ich kann mich immer auf eine künstliche Intelligenz verlassen, wenn es um die Entscheidungsfindung geht.
12. Ich kann den Informationen vertrauen, die mir eine künstliche Intelligenz liefert.

NASA Task Load Index (NASA-TLX)

1. Wie viel geistige Anforderung war bei der Aufnahme und Verarbeitung von Informationen erforderlich? War die Aufgabe einfach oder komplex?
2. Wie erfolgreich haben Sie Ihrer Meinung nach die vom Versuchsleiter (oder Ihnen selbst) gesetzten Ziele erreicht?
3. Wie anstrengend war die Arbeit, um Ihren Grad an Aufgabenerfüllung zu erreichen?
4. Wie frustriert (unsicher, entmutigt, irritiert, gestresst und verärgert) fühlten Sie sich während der Aufgabe?

Ethics Position Questionnaire (EPQ)

Bitte geben Sie den Grad Ihrer Zustimmung zu folgenden Aussagen an.

1. Das Wohl anderer zu opfern, ist niemals wirklich notwendig.
2. Moralische Standards sollten als etwas Individuelles gesehen werden: Was eine Person als moralisch ansieht, kann eine andere als unmoralisch bewerten.
3. Die Würde und das Wohlergehen der Menschen sollten die wichtigste Sorge in jeder Gesellschaft sein.
4. Ob eine Lüge als unmoralisch oder sogar moralisch zu beurteilen ist, hängt ganz von den Umständen ab.
5. In sozialen Beziehungen sind ethische Probleme oft so komplex, dass man Personen erlauben sollte, ihre eigenen persönlichen Regeln zu finden.
6. Was „ethisch“ ist, variiert zwischen Situationen und Kulturen.
7. Es ist unmoralisch, negative Folgen einer Handlung durch positive Folgen verrechnen zu wollen.
8. Man darf andere Personen weder psychisch noch physisch schädigen.
9. Wenn eine Handlung eine unschuldige Person schädigen könnte, muss man sie unterlassen.
10. Es gibt keine ethischen Prinzipien, die so wichtig sind, dass sie eine allgemeingültige Vorschrift bilden könnten.
11. Moralisches Handeln liegt dann vor, wenn es der Ideal-Handlung entspricht.
12. Man darf keine Handlungen ausführen, die in irgendeiner Weise die Würde und das Wohlergehen anderer Personen bedrohen.
13. Eine starre Ethik-Vorschrift, die bestimmte Handlungsmöglichkeiten verhindern soll, kann der Verbesserung sozialer Beziehungen sogar im Wege stehen.
14. Risiken in Kauf zu nehmen, die andere Personen betreffen, ist nicht tolerierbar, egal wie gering sie sind.
15. Potentielle Schädigungen Dritter in Kauf zu nehmen, ist immer schlecht, egal welche guten Zwecke verfolgt werden.
16. Moralisches Standards sind jeweils persönliche Regeln, sie sollten nicht auf die Beurteilung anderer angewendet werden.
17. Die Frage, was ethisch richtig ist, wird sich niemals beantworten lassen, da es sich bei der Entscheidung, was moralisch oder unmoralisch ist, um eine persönliche Entscheidung handelt.
18. Man sollte sichergehen, mit seinen Handlungen niemanden zu verletzen oder zu schädigen.
19. Verschiedene Arten von Moral dürfen nicht als mehr oder weniger „Gut“ bewertet werden.
20. Über das Lügen lässt sich keine Regel formulieren; ob eine Lüge zulässig ist oder nicht, hängt von der Situation ab.

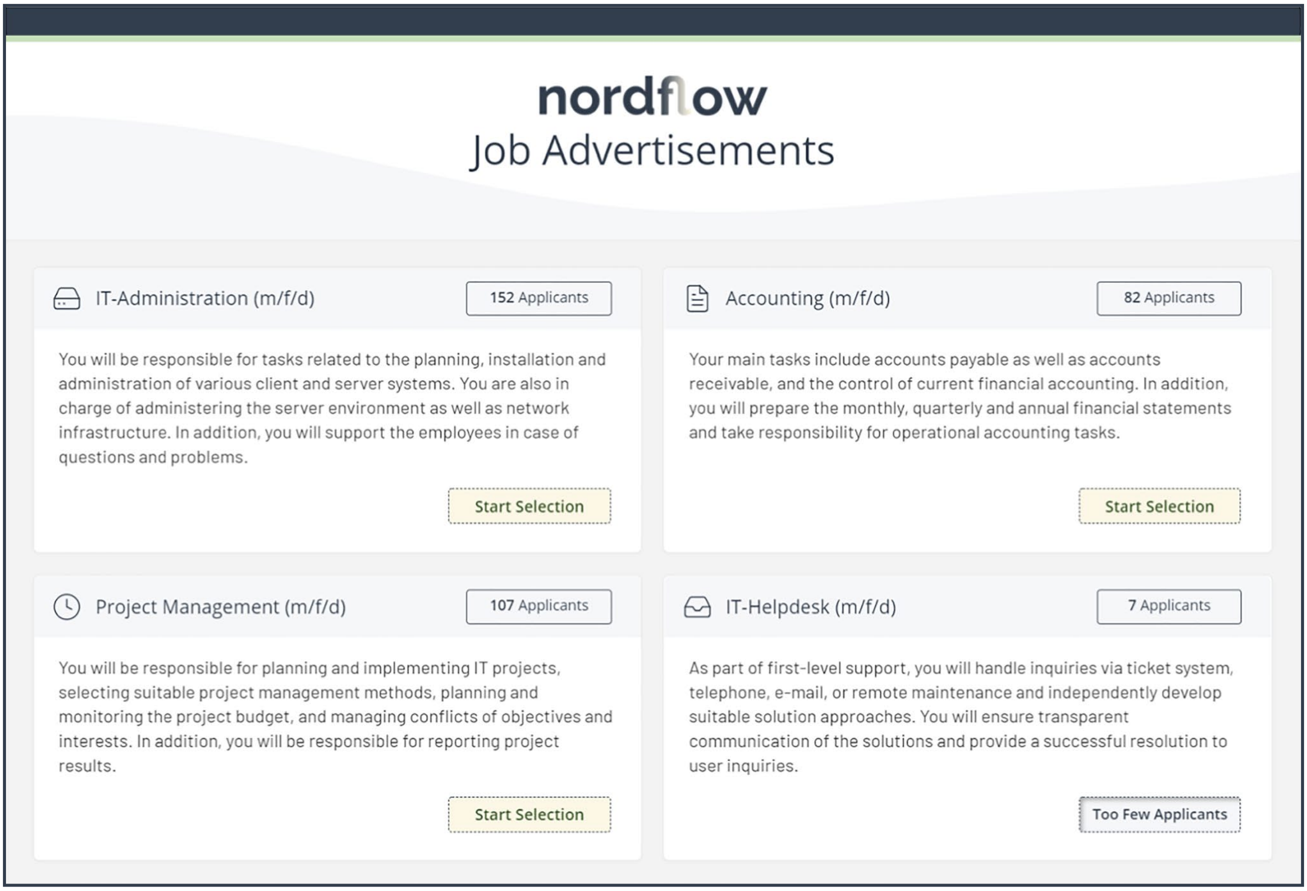
Appendix 4

Information about the AI-based system (translated from German)

Important note on the implemented AI technology

The implemented AI makes use of various algorithms. During development, a great focus was placed on fair and ethical decisions. The AI distinguishes between applicants on the basis of a variety of features (characteristics) that have been selected by an independent panel of experts. Access to the applicants' personal information is technically prevented. The measures implemented ensure that the evaluation carried out by the AI is always based on objective factors. The aim is to encourage the decision-maker to make more ethical decisions in candidate selection processes.

Appendix 5



nordflow

Job Advertisements

IT-Administration (m/f/d) — 152 Applicants

You will be responsible for tasks related to the planning, installation and administration of various client and server systems. You are also in charge of administering the server environment as well as network infrastructure. In addition, you will support the employees in case of questions and problems.

Accounting (m/f/d) — 82 Applicants

Your main tasks include accounts payable as well as accounts receivable, and the control of current financial accounting. In addition, you will prepare the monthly, quarterly and annual financial statements and take responsibility for operational accounting tasks.

Project Management (m/f/d) — 107 Applicants

You will be responsible for planning and implementing IT projects, selecting suitable project management methods, planning and monitoring the project budget, and managing conflicts of objectives and interests. In addition, you will be responsible for reporting project results.

IT-Helpdesk (m/f/d) — 7 Applicants

As part of first-level support, you will handle inquiries via ticket system, telephone, e-mail, or remote maintenance and independently develop suitable solution approaches. You will ensure transparent communication of the solutions and provide a successful resolution to user inquiries.

Fig. 7 “Job view” of the platform displaying job advertisements

Job descriptions

BFI-10 findings

The BFI-10 questionnaire was assessed on a 7-point Likert scale to evaluate the participants' personalities. To calculate the BFI scores, the negatively formulated items were inverted, and the mean values of the two items of each of the five dimensions were calculated. Afterward, descriptive statistics were generated, as shown in the Table 8 below. The BFI-scores of the groups with and without AI recommendations and the BFI scores of the groups with and without XAI were compared with independent samples t-tests. There were no significant differences between the respective groups.

Table 8 Results of BFI-10

Dimension	$M\ (SD)_{AI\ rec}$	$M\ (SD)_{No\ rec}$	$M\ (SD)_{XAI}$	$M\ (SD)_{No\ XAI}$	$M\ (SD)_{Total}$
Extraversion	4.09 *(1.5)*	4.19 *(1.50)*	4.19 *(1.44)*	4.09 *(1.58)*	4.14 *(1.50)*
Agreeableness	4.27 *(1.3)*	4.38 *(1.16)*	4.23 *(1.29)*	4.42 *(1.17)*	4.33 *(1.23)*
Conscientiousness	4.95 *(1.26)*	5.12 *(1.15)*	5.06 *(1.13)*	5.00 *(1.29)*	5.03 *(1.21)*
Neuroticism	4.18 *(1.38)*	4.22 *(1.3)*	4.12 *(1.33)*	4.27 *(1.35)*	4.20 *(1.34)*
Openness	5.05 *(1.39)*	4.93 *(1.54)*	5.08 *(1.49)*	4.90 *(1.44)*	4.99 *(1.46)*

Appendix 6

Participants reasons

https://osf.io/3tjre/?view_only=47915ebaca684083acd568b9a7b4941b

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12525-022-00600-9.

Author contributions Lennart Hofeditz and Stefan Stieglitz contributed to the study conception and design. Material preparation, data collection and analysis were performed by Alexander Rieß and Lennart Hofeditz. The first draft of the manuscript was written by Lennart Hofeditz and Sünje Clausen. Milad Mirbabaie and Stefan Stieglitz supervised the research project, provided senior scholar knowledge, and helped revising the manuscript for submission. All authors commented on previous versions of the manuscript and read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest Thus, there are no financial or other conflicting interests in this work.

References

Abrams, D., Swift, H. J., & Drury, L. (2016). Old and unemployable? How age-based stereotypes affect willingness to hire job candidates. *Journal of Social Issues, 72*(1), 105–121. https://doi.org/10.1111/josi.12158

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Akinlade, E. Y., Lambert, J. R., & Zhang, P. (2020). Mechanisms for hiring discrimination of immigrant applicants in the United States. *Equality, Diversity and Inclusion: An International Journal, 39*(4), 395–417. https://doi.org/10.1108/EDI-08-2019-0218

Ameri, M., Schur, L., Adya, M., Bentley, F. S., McKay, P., & Kruse, D. (2018). The disability employment puzzle: A field experiment on employer hiring behavior. *ILR Review, 71*(2), 329–364. https://doi.org/10.1177/0019793917717474

Baert, S. (2018). Hiring discrimination: An overview of (almost) all correspondence experiments since 2005. In *Audit studies: Behind the scenes with theory, method, and nuance* (pp. 63–77). Springer International Publishing. https://doi.org/10.1007/978-3-319-71153-9_3

Baert, S., Albanese, A., du Gardein, S., Ovaere, J., & Stappers, J. (2017). Does work experience mitigate discrimination? *Economics Letters, 155*(July 2013), 35–38. https://doi.org/10.1016/j.econlet.2017.03.011

Baird, A., & Maruping, L. M. (2021). The next generation of research on is use: A theoretical framework of delegation to and from agentic is artifacts. *MIS Quarterly Management Information Systems, 45*(1), 315–341. https://doi.org/10.25300/MISQ/2021/15882

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *SSRN Electronic Journal, 104*(671), 671–732. https://doi.org/10.2139/ssrn.2477899

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*(December 2019), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly, 45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business and Information Systems Engineering, 63*(1), 55–68. https://doi.org/10.1007/s12599-020-00678-5

Bigman, Y. E., Yam, K. C., Marciano, D., Reynolds, S. J., & Gray, K. (2021). Threat of racial and economic inequality increases preference for algorithm decision-making. *Computers in Human Behavior, 122*(April), 106859. https://doi.org/10.1016/j.chb.2021.106859

Black, J. S., & van Esch, P. (2020). AI-enabled recruiting: What is it and how should a manager use it? *Business Horizons, 63*(2), 215–226. https://doi.org/10.1016/j.bushor.2019.12.001

Burke, G., Mendoza, M., Linderman, J., & Tarm, M. (2021). *How AI-powered tech landed man in jail with scant evidence*. Associated Press.

Carlsson, R., & Sinclair, S. (2018). Prototypes and same-gender bias in perceptions of hiring discrimination. *The Journal of Social Psychology, 158*(3), 285–297. https://doi.org/10.1080/00224545.2017.1341374

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior, 127*, 107018. https://doi.org/10.1016/j.chb.2021.107018

Cole, M. S., Feild, H. S., & Giles, W. F. (2004). Interaction of recruiter and applicant gender in resume evaluation: A field study. *Sex Roles, 51*(9–10), 597–608. https://doi.org/10.1007/s11199-004-5469-1

Correll, S. J., Benard, S., & Paik, I. (2007). Getting a job: Is there a motherhood penalty? *American Journal of Sociology, 112*(5), 1297–1339. https://doi.org/10.1086/511799

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Davidson, S. (2016). Gender inequality: Nonbinary transgender people in the workplace. *Cogent Social Sciences, 2*(1), 1236511. https://doi.org/10.1080/23311886.2016.1236511

Davidson, R., & MacKinnon, J. (1993). *Estimation and inference in econometrics*. Oxford University Press.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business and Information Systems Engineering, 61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human Computer Studies, 162*(September 2021), 102792. https://doi.org/10.1016/j.ijhcs.2022.102792

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L. C., Misra, S., Mogaji, E., Kumar Sharma, S., Bahadur Singh, J., Raghavan, V., Raman, R., Rana, N. P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji A., Waltony, P., & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Ebel, P., Söllner, M., Leimeister, J. M., Crowston, K., & de Vreede, G.-J. (2021). Hybrid intelligence in business networks. *Electronic Markets, 31*(2), 313–318. https://doi.org/10.1007/s12525-021-00481-4

Feloni, R. (2017). *Consumer goods giant Unilever has been hiring employees using brain games and artificial intelligence — and it's a huge success*. Business Insider Australia. https://www.businessinsider.in/Consumer-goods-giant-Unilever-has-been-hiring-employees-using-brain-games-and-artificial-intelligence-and-its-a-huge-success/articleshow/59356757.cms

Fernández-Martínez, C., & Fernández, A. (2020). AI and recruiting software: Ethical and legal implications. *Paladyn, Journal of Behavioral Robotics, 11*(1), 199–216. https://doi.org/10.1515/pjbr-2020-0030

Fiske, S. T., Bersoff, D. N., Borgida, E., Deaux, K., & Heilman, M. (1991). Social science research on trial: Use of sex stereotyping research in Price Waterhouse v. Hopkins. *American Psychologist, 46*(10), 1049–1060. https://doi.org/10.1037/0003-066X.46.10.1049

Foley, M., & Williamson, S. (2018). Does anonymising job applications reduce gender bias? *Gender in Management: An International Journal, 33*(8), 623–635. https://doi.org/10.1108/GM-03-2018-0037

Foschi, M., Lai, L., & Sigerson, K. (1994). Gender and double standards in the assessment of job applicants. *Social Psychology Quarterly, 57*(4), 326. https://doi.org/10.2307/2787159

Franke, T., Attig, C., & Wessel, D. (2017). *Assessing affinity for technology interaction – the affinity for technology assessing affinity for technology interaction (ATI). July*. https://doi.org/10.13140/RG.2.2.28679.50081

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 80–89). https://doi.org/10.1109/DSAA.2018.00018

González, M. J., Cortina, C., & Rodríguez, J. (2019). The role of gender stereotypes in hiring: A field experiment. *European Sociological Review, 35*(2), 187–204. https://doi.org/10.1093/esr/jcy055

Gorges, H. (2015). *HR braucht mehr Männer*. Human Resources Manager. https://www.humanresourcesmanager.de/recruiting/hr-braucht-mehr-maenner/

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42. https://doi.org/10.1145/3236009

Gulati, S. N., Sousa, S. C., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour and Information Technology, 38*(10), 1004–1015. https://doi.org/10.1080/0144929X.2019.1656779

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Guryan, J., & Charles, K. K. (2013). Taste-based or statistical discrimination: The economics of discrimination returns to its roots. *The Economic Journal, 123*(572), F417–F432. https://doi.org/10.1111/ecoj.12080

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Power Technology and Engineering, 43*(5), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based perspective* (2nd ed.). Guilford Press.

Hepenstal, S., & McNeish, D. (2020). Explainable artificial intelligence: What do you need to know? In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): Vol. 12196 LNAI* (Issue Lipton 2016). Springer International Publishing. https://doi.org/10.1007/978-3-030-50353-6_20

Hofeditz, L., Mirbabaie, Mi., Stieglitz, S., & Holstein, J. (2021). Do you trust an AI-Journalist? A credibility analysis of news content with AI-Authorship. *Proceedings of the 28th European Conference on Information Systems*. Marakech, Morocco.

Hofeditz, L., Harbring, M., Mirbabaie, M., & Stieglitz, S. (2022a). Working with ELSA – how an emotional support agent builds trust in virtual teams. *Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii.

Hofeditz, L., Mirbabaie, M., Luther, A., Mauth, R., & Rentemeister, I. (2022b). Ethics guidelines for using ai-based algorithms in recruiting: Learnings from a systematic literature review. *Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Metrics for explainable AI: Challenges and prospects* (pp. 1–50).https://doi.org/10.48550/arXiv.1812.04608

Houser, K. A. (2019). Can AI solve the diversity problem in the tech industry? Mitigating noise and bias in employment decision-making. *Stanford Technology Law Review, 22*(2), 291–353.

Hu, J. (2019). *99% of Fortune 500 Companies use Applicant Tracking Systems*. Jobscan. https://www.jobscan.co/blog/99-percent-fortune-500-ats/

Hussain, F., Hussain, R., & Hossain, E. (2021). Explainable Artificial Intelligence (XAI): An engineering perspective. *arXiv*, 1–11.https://doi.org/10.48550/arXiv.2101.03613

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research, 2nd ed.* (pp. 102–138). Guilford Press.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *Proceedings of the 28th European Conference on Information Systems.* Marakech, Morocco.

Köchling, A., Riazy, S., Wehner, M. C., & Simbeck, K. (2021). Highly accurate, but still discriminatory: A fairness evaluation of algorithmic video analysis in the recruitment context. *Business and Information Systems Engineering, 63*(1), 39–54. https://doi.org/10.1007/s12599-020-00673-w

Krause, A., Rinne, U., & Zimmermann, K. F. (2012). Anonymous job applications in Europe. *IZA Journal of European Labor Studies, 1*(1), 5. https://doi.org/10.1186/2193-9012-1-5

Kübler, D., Schmid, J., & Stüber, R. (2018). Gender discrimination in hiring across occupations: A nationally-representative vignette study. *Labour Economics, 55*, 215–229. https://doi.org/10.1016/j.labeco.2018.10.002

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2019). Search bias quantification: Investigating political bias in social media and web search. *Information Retrieval Journal, 22*(1–2), 188–227. https://doi.org/10.1007/s10791-018-9341-2

Kuncel, N. R., Klieger, D. M., & Ones, D. S. (2014). In hiring, algorithms beat instinct. *Harvard Business Review, 92*, 32.

Lancee, B. (2021). Ethnic discrimination in hiring: comparing groups across contexts. Results from a cross-national field experiment. *Journal of Ethnic and Migration Studies, 47*(6), 1181–1200. https://doi.org/10.1080/1369183X.2019.1622744

Laurim, V., Arpaci, S., Prommegger, B., & Krcmar, H. (2021). Computer, whom should I hire? - Acceptance criteria for artificial intelligence in the recruitment process. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2020-Janua* (pp. 5495–5504). https://doi.org/10.24251/hicss.2021.668

Lepak, D. P., Marrone, J. A., & Takeuchi, R. (2004). The relativity of HR systems: Conceptualising the impact of desired employee contributions and HR philosophy. *International Journal of Technology Management, 27*(6–7), 639–655. https://doi.org/10.1504/IJTM.2004.004907

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology, 31*(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

Li, L., Lassiter, T., Oh, J., & Lee, M. K. (2021). Algorithmic hiring in practice: Recruiter and HR professional's perspectives on AI use in hiring. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, *1*(1), 166–176. https://doi.org/10.1145/3461702.3462531

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: informing design practices for explainable AI user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). https://doi.org/10.1145/3313831.3376590

Lössbroek, J., Lancee, B., van der Lippe, T., & Schippers, J. (2021). Age discrimination in hiring decisions: A factorial survey among managers in nine European countries. *European Sociological Review, 37*(1), 49–66. https://doi.org/10.1093/esr/jcaa030

Mayring, P. (1994). *Qualitative Inhaltsanalyse*. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-14565

Mehrotra, A., & Celis, L. E. (2021). Mitigating bias in set selection with noisy protected attributes. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 237–248). https://doi.org/10.1145/3442188.3445887

Meske, C., & Bunde, E. (2020). Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support. In H. Degen & L. Reinerman-Jones (Eds.), *Artificial intelligence in HCI* (pp. 54–69). Springer International Publishing.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management, 39*(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Mingers, J., & Walsham, G. (2010). Toward ethical information systems: The contribution of discourse ethics. *MIS Quarterly, 34*(4), 833–854. https://doi.org/10.2307/25750707

Mirbabaie, M., Brünker, F., Möllmann (Frick), N. R. J., & Stieglitz, S. (2022). The rise of artificial intelligence – understanding the AI identity threat at the workplace. *Electronic Markets, 32*(1), 73–99. https://doi.org/10.1007/s12525-021-00496-x

Mirbabaie, M., Stieglitz, S., Brünker, F., Hofeditz, L., Ross, B., & Frick, N. R. J. (2021a). Understanding collaboration with virtual assistants – the role of social identity and the extended self. *Business and Information Systems Engineering, 63*(1), 21–37. https://doi.org/10.1007/s12599-020-00672-x

Mirbabaie, M., Stieglitz, S., & Frick, N. R. J. (2021b). Hybrid intelligence in hospitals: Towards a research agenda for collaboration. *Electronic Markets, 31*(2) 365–387. https://doi.org/10.1007/s12525-021-00457-4

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 279–288). https://doi.org/10.1145/3287560.3287574

Mujtaba, D. F., & Mahapatra, N. R. (2019). Ethical considerations in AI-based recruitment. *2019 IEEE International Symposium on Technology and Society (ISTAS)* (pp. 1–7). https://doi.org/10.1109/ISTAS48451.2019.8937920

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature, 56*(3), 799–866. https://doi.org/10.1257/jel.20161309

Neumark, D. (2021). Age discrimination in hiring: Evidence from age-blind vs. non-age-blind hiring procedures. *Journal of Human Resources*, *August*, 0420-10831R1. https://doi.org/10.3368/jhr.0420-10831R1

Neumark, D., Burn, I., & Button, P. (2017). Age discrimination and hiring of older workers. *FRBSF Economic Letter, 06*(2014), 1–5.

Ochmann, J., & Laumer, S. (2019). Fairness as a determinant of AI adoption in recruiting: An interview-based study. *DIGIT 2019 Proceedings*. https://aisel.aisnet.org/digit2019/16

Ochmann, J., Zilker, S., Michels, L., Tiefenbeck, V., & Laumer, S. (2021). The influence of algorithm aversion and anthropomorphic agent design on the acceptance of AI-based job recommendations. *International Conference on Information Systems, ICIS 2020* (pp. 17).

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Pan, Y., Froese, F., Liu, N., Hu, Y., & Ye, M. (2021). The adoption of artificial intelligence in employee recruitment: The influence of contextual factors. *The International Journal of Human Resource Management*, 1–23. https://doi.org/10.1080/09585192.2021.1879206

Petersen, T., & Saporta, I. (2004). The opportunity structure for discrimination. *American Journal of Sociology, 109*(4), 852–901. https://doi.org/10.1086/378536

Petersen, T., & Togstad, T. (2006). Getting the offer: Sex discrimination in hiring. *Research in Social Stratification and Mobility, 24*(3), 239–257. https://doi.org/10.1016/j.rssm.2006.06.001

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences of the United States of America, 114*(41), 10870–10875. https://doi.org/10.1073/pnas.1706255114

Quillian, L., Heath, A., Pager, D., Midtbøen, A., Fleischmann, F., & Hexel, O. (2019). Do some countries discriminate more than others? Evidence from 97 field experiments of racial discrimination in hiring. *Sociological Science, 6*, 467–496. https://doi.org/10.15195/v6.a18

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 469–481). https://doi.org/10.1145/3351095.3372828

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review, 46*(1), 192–210. https://doi.org/10.5465/AMR.2018.0072

Rammstedt, B., Kemper, C., Klein, M., Beierlein, C., & Kovaleva, A. (2013). Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit: 10 Item Big Five Inventory (BFI-10). *Methoden, Daten, Analysen (Mda)*, *7*(2), 233–249. https://doi.org/10.12758/mda.2013.013

Richardson, B., Webb, J., Webber, L., & Smith, K. (2013). *Age discrimination in the evaluation of job applicants: Discovery Service for University of Portsmouth* (pp. 35–44). https://doi.org/10.1111/j.1559-1816.2013.00979.x

Rieskamp, J., Hofeditz, L., Mirbabaie, M., & Stieglitz, S. (2023). Approaches to improve fairness when deploying ai-based algorithms in hiring – using a systematic literature review to guide future research. *Hawaii International Conference on System Sciences*. Maui, Hawaii.

Rouse, W. B. (2020). AI as systems engineering augmented intelligence for systems engineers. *Insight, 23*(1), 52–54. https://doi.org/10.1002/inst.12286

Ruffle, B. J., & Shtudiner, Z. (2015). Are good-looking people more employable? *Management Science, 61*(8), 1760–1776. https://doi.org/10.1287/mnsc.2014.1927

Russell, S. J., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Pearson Education Limited.

Sabeg, Y., & Me´haignerie, L. (2006). *Les oublie´s de l'e´galite´ des chances [The forgotten ones of the equality of opportunity]*. Hachette.

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to "solve" the problem of discrimination in hiring? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 458–468). https://doi.org/10.1145/3351095.3372849

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems, 29*(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094

Schoonderwoerd, T. A. J., Zoelen, E. M. va., Bosch, K. van den, & Neerincx, M. A. (2022). Design patterns for human-AI co-learning: A wizard-of-Oz evaluation in an urban-search-and-rescue task. *International Journal of Human Computer Studies, 164*(July 2021), 102831. https://doi.org/10.1016/j.ijhcs.2022.102831

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Sokol, K., & Flach, P. (2020). Explainability fact sheets. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 56–67). https://doi.org/10.1145/3351095.3372870

Stone, A., & Wright, T. (2013). When your face doesn't fit: Employment discrimination against people with facial disfigurements. *Journal of Applied Social Psychology, 43*(3), 515–526. https://doi.org/10.1111/j.1559-1816.2013.01032.x

Strack, M., & Gennerich, C. (2007). Erfahrung mit Forsyths 'Ethic Position Questionnaire? (EPQ): Bedeutungsunabhängigkeit von Idealismus und Realismus oder Akquieszens und Biplorarität? *Berichte Aus Der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral", Nr. 167, ISSN 1430-1148*.

Sühr, T., Hilgard, S., & Lakkaraju, H. (2021). Does fair ranking improve minority outcomes? Understanding the interplay of human and algorithmic biases in online hiring. *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 989–999). https://doi.org/10.1145/3461702.3462602

Teodorescu, M. H. M., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of fairness in automation require a deeper understanding of human–ml augmentation. *MIS Quarterly: Management Information Systems, 45*(3), 1483–1499. https://doi.org/10.25300/MISQ/2021/16535

Thiebes, S., Lins, S., & Sunyaev, A. (2020). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Tosi, H. L., & Einbender, S. W. (1985). The effects of the type and amount of information in sex discrimination research: A meta-analysis. *Academy of Management Journal, 28*(3), 712–723. https://doi.org/10.5465/256127 van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research, 144*(February), 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

Weichselbaumer, D. (2016). Discrimination against female migrants wearing headscarves. *SSRN Electronic Journal*, *10217*. https://doi.org/10.2139/ssrn.2842960

Weiss, A., Bernhaupt, R., Schwaiger, D., Altmaninger, M., Buchner, R., & Tscheligi, M. (2009). User experience evaluation with a Wizard of Oz approach: Technical and methodological considerations. *9th IEEE-RAS International Conference on Humanoid Robots*, (pp. 303–308). https://doi.org/10.1109/ICHR.2009.5379559

Wijnhoven, F. (2021). Organizational learning for intelligence amplification adoption: Lessons from a clinical decision support system adoption project. *Information Systems Frontiers*, *0123456789*.https://doi.org/10.1007/s10796-021-10206-9

Wijnhoven, F., & van Haren, J. (2021). Search engine gender bias. *Frontiers in Big Data, 4*(May), 1–12. https://doi.org/10.3389/fdata.2021.622106

Wilson, J., & Rosenberg, D. (1988). Rapid prototyping for user interface design. In *Handbook of human-computer interaction*. Elsevier B.V. https://doi.org/10.1016/b978-0-444-70536-5.50044-0

Zaniboni, S., Kmicinska, M., Truxillo, D. M., Kahn, K., Paladino, M. P., & Fraccaroli, F. (2019). Will you still hire me when I am over 50? The effects of implicit and explicit age stereotyping on resume evaluations. *European Journal of Work and Organizational Psychology, 28*(4), 453–467. https://doi.org/10.1080/1359432X.2019.1600506

Zhu, J., Liapis, A., Risi, S., Bidarra, R., & Youngblood, G. M. (2018). Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, *2018-Augus* (pp. 1–8). https://doi.org/10.1109/CIG.2018.8490433

Zschirnt, E., & Ruedin, D. (2016). Ethnic discrimination in hiring decisions: A meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies, 42*(7), 1115–1134. https://doi.org/10.1080/1369183X.2015.1133279

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paper 2: Ethics and AI in Information Systems Research

Type (Ranking, Impact Factor)	Journal article (C, 2.38)
Status	Published
Rights and permissions	Open access
Authors	Mirbabaie, M., Brendel, A. B., & Hofeditz, L.
Year	2022
Outlet	Communications of the Association for Information Systems (CAIS)
Permalink / DOI	https://doi.org/10.17705/1CAIS.05034
Full citation	Mirbabaie, M., Brendel, A. B., & Hofeditz, L. (2022). Ethics and AI in Information Systems Research. *Communication of the Association for Information Systems*, 50, 123–150. https://doi.org/10.17705/1CAIS.05034.

Communications of the Association for Information Systems

Volume 50 Article 38

6-27-2022

Ethics and AI in Information Systems Research

Milad Mirbabaie
Paderborn University, milad.mirbabaie@uni-paderborn.de

Alfred B. Brendel
Faculty of Business and Economics, Technische Universität Dresden

Lennart Hofeditz
Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen

Recommended Citation

Mirbabaie, M., Brendel, A. B., & Hofeditz, L. (2022). Ethics and AI in Information Systems Research. Communications of the Association for Information Systems, 50, pp-pp. https://doi.org/10.17705/1CAIS.05034

Communications of the Association for Information Systems

Special Section DOI: 10.17705/1CAIS.05034 ISSN: 1529-3181

Ethics and AI in Information Systems Research

Milad Mirbabaie
Department of Information Systems, Paderborn University
milad.mirbabaie@uni-paderborn.de

Alfred Benedikt Brendel
Faculty of Business and Economics, Technische Universität Dresden

Lennart Hofeditz
Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen

Abstract:

The ethical dimensions of Artificial Intelligence (AI) constitute a salient topic in information systems (IS) research and beyond. There is an increasing number of journal and conference articles on how AI should be designed and used. For this, IS research offers and curates knowledge not only on the ethical dimensions of information technologies but also on their acceptance and impact. However, the current discourse on the ethical dimensions of AI is highly unstructured and seeks clarity. As conventional systematic literature research has been criticized for lacking in performance, we applied an adapted discourse approach to identify the most relevant articles within the debate. As the fundamental manuscripts within the discourse were not obvious, we used a weighted citation-based technique to identify fundamental manuscripts and their relationships within the field of AI ethics across disciplines. Starting from an initial sample of 175 papers, we extracted and further analyzed 12 fundamental manuscripts and their citations. Although we found many similarities between traditionally curated ethical principles and the identified ethical dimensions of AI, no IS paper could be classified as fundamental to the discourse. Therefore, we derived our own ethical dimensions on AI and provided guidance for future IS research.

Keywords: Ethics, Artificial Intelligence, Discourse Approach, Review Article, Information Systems.

This manuscript underwent peer review. It was received 9/1/2020 and was with the authors for 18 months for three revisions. The Associate Editor chose to remain anonymous.

1 Introduction

While organizations and researchers have repeatedly shown the advantages of Artificial Intelligence (AI)-based systems for humanity (such as self-driving cars, AI-based conversational agents, and process automation), serious AI-related abuses and incidents have raised pressing ethical concerns (Benbya et al., 2021; Berente et al., 2021; Seppälä et al., 2021). While unethical behavior can be intended in some cases due to skewed organizational or managerial values (e.g. during the VW diesel scandal) (Stieglitz et al., 2019), many unintended ethical challenges and moral issues can occur when applying AI (Boddington, 2017). For instance, Amazon's discriminatory human resources (HR) software and Microsoft's racist chatbot provide a strong case for the dangerous and unethical sides of AI that were inadvertent (Dastin, 2018; Horton, 2016; Yampolskiy, 2016). Furthermore, organizations such as Uber increasingly apply AI-based algorithms for exerting autonomous managerial control over employees (referred to as algorithmic control), resulting in constant surveillance, less transparency and possible dehumanization (Wiener et al., 2021). On the one hand, these unethical sides of AI and algorithms are grounded in biased man-made algorithms. The latter are used, for instance, in hiring, and cannot be completely non-discriminatory (Mann & O'Neil, 2016). On the other hand, this is due to the predictive nature of AI, resulting in a non-transparent derivation of outputs (Boddington, 2017).

There have been many different approaches to defining artificial intelligence in the past. Carvalho et al. (2019) considered AI as a group of technologies that rely on techniques such as machine learning, natural language processing, and knowledge representation. However, we do not consider AI to be a single technology or a group of specific technologies. We follow the definition of an AI as *"the frontier of computational advancements that references human intelligence in addressing ever more complex decision-making problems" (Berente et al., 2021)*.

There is, however, a conflict between AI and ethics. Advances in AI technologies require increasing amounts of data to make AI work properly while, at the same time, the technology is being given more and more autonomy. Normative ethics, in contrast, aim to protect the rights of individuals, including data and autonomy. AI technologies can be applied to many different use cases. It is, therefore, difficult for organizations, researchers, and policymakers to draw up ethical guidelines that are neither too narrowly targeted on a specific use case, nor too vague. In addition, organizations such as IBM have defined ethical principles for themselves, although this does not prevent them from pursuing unethical AI activities (Robin, 2019). Researchers will need to address this conflict between ethics and AI and develop strategies to resolve it.

Accordingly, the increasing influence of AI on society as well as individuals goes hand-in-hand with the increasing pressure on organizations to acknowledge responsibility for their AI products and offerings (Brendel et al., 2021). This includes ethical considerations as to their AI's potential consequences. For leaders to incorporate ethical considerations into their decisions when applying AI in their organizations, they need guidance from research. With knowledge in both normative ethics (e.g., Stahl, 2012) and organizational processes, the information systems (IS) community clearly offers the potential to take an interdisciplinary bridging role in the ethical application of AI-based systems. Previous research proved that examining the effects of digitalization on principles such as human dignity is an area in which IS scholars can contribute valuable artifacts (Leidner & Tona, 2021). IS researchers connect knowledge from different disciplines and provide theories that can be used to understand and interpret emerging phenomena such as AI. An ethical discourse on AI that has been widely acknowledged by researchers from different disciplines lacks such an interdisciplinary link that IS research can provide (Brendel et al., 2021). Research on AI ethics resides within multiple domains, including but not limited to philosophy, IS, computer sciences, and social or management research (Bostrom & Yudkowsky, 2014). The renunciation of this ethical discourse, which by its philosophical and multidimensional nature tends to be controversial, can entail significant and considerable consequences and risks for our society (Boddington, 2017). With AI-based systems, it is, therefore, important to create guidelines for dealing with AI at an early stage.

In the last decade, IS researchers have focused on designing new artifacts, particularly AI applications (Ahsen et al., 2019; Kloör et al., 2018) or on examining AI applications in certain application domains such as healthcare (Mirbabaie et al., 2021a) or media distribution (Hofeditz et al., 2021). However, as AI becomes increasingly more capable, only focusing on AI's positive side can be misleading or even dangerous. Therefore, some IS scholars have begun to establish a discourse related to the ethical challenges of AI (e.g., Mendling et al., 2018; Porra et al., 2019). Primary examples of ethical considerations include the greater complexity of AI and its increasing decision-making autonomy. The complexity makes it harder to understand how and why an AI has come to a certain decision—and what decisions it will make in the future (Gunning, 2017). The increasing decision-making autonomy of AI concerns decisions that an AI is able to take on its own with little or no prior human approval or supervision (Kalenka & Jennings, 1999). A prominent concept in this context is algorithmic aversion (Berger et al., 2021; Kawaguchi, 2021; Renier et al., 2021). This phenomenon, which has been illustrated by various studies (e.g., Dietvorst et al., 2015; Dietvorst et al., 2018), shows that human decision makers tend to consider algorithmic forecasters significantly less than human forecasters, even if the humans repeatedly perform worse in the forecasts. Furthermore, decision makers tend to choose a modifiable imperfect algorithm over a non-modifiable perfect algorithm. One reason for this algorithmic aversion is the desire of individuals to have at least some level of control and autonomy (Dietvorst et al., 2018). However, this possibility for autonomy and indivisibility is not present in every AI-based system. There are also studies showing that laypeople are more likely to trust algorithms than humans for certain predictions, which can be called algorithmic appreciation (Logg et al., 2019). This shows that algorithmic aversion is not always as straightforward as it might seem, and that future research needs to look further into this and other related ethical dimensions and phenomena. Another ethically relevant area in the context of AI-based systems is trust in the system (Hofeditz et al., 2021; Mirbabaie et al., 2021a; Thiebes et al., 2021). One recently published study suggests that a loss of trust in familiar AI-based systems due to perceived errors of a familiar system over time is one possible explanation of algorithmic aversion (Berger et al., 2021).

The discourse on ethical dimensions of AI with the potential of IS to assume a leading role due to its interdisciplinary knowledge seems to be highly unstructured, and we hardly found any established theory papers in this field. We found some promising conference pieces dealing with the implementation of AI ethics in organizations (Mayer et al., 2021), a governance framework for AI regulation (de Almeida et al., 2020), and ethical implications of bias in machine learning (Yapo & Weiss, 2018). However, with an initial search, we neither found often-cited high-quality IS journal publications nor articles providing guidance for IS research on how to systematically examine the ethical dimensions of AI. In addition, some domains, such as healthcare or quality management for materials, could, from an ethical point of view, be considered more important than others. In sensible cases, ethical discourse must be discussed more compellingly compared to less sensitive cases. However, the IS discourse has not elaborated on that so far. To the best of our knowledge, the individual conclusions on the ethical dimensions and implications of AI reside within various domains, hiding a common foundation of what is known and what needs to be addressed in practice and research. The foundations of the ethical dimensions of AI seem to be widely scattered and ambiguous. Therefore, we ask the following research question:

> **RQ: What is the status quo of IS research regarding the discourse on the ethical dimensions of AI?**

Against this background, we aim to gather research from various sources, extending beyond the scope of the AIS basket of eight journals and prominent IS conferences (e.g., ICIS, ECIS, PACIS, AMCIS). To contribute to the discourse with knowledge of the ethical dimensions of AI, we identified and analyzed the domain ecosystem via the application of a discourse approach to corpus construction (Larsen et al., 2019), including consecutive forward and backward searches. Starting from an IS perspective, but also including various works from outside IS in the backward and forward search, we gathered 125 relevant papers from several disciplines and identified 12 fundamental manuscripts on the ethical dimensions of AI. By analyzing the gathered literature, we identified research gaps within the ecosystem and derived directions for IS research. With our review, we aim to provide a base for future research directions on the ethical dimensions of AI inside the IS community, hopefully jumpstarting a rich exchange between disciplines.

This paper is structured as follows. In Section 2, we highlight the importance of ethics in IS research and outline the current state of research on the ethical dimensions of AI. In Section 3, we describe why and how we used the discourse approach, according to Larsen et al. (2019). In Section 4, we summarize our findings and provide an overview of the fundamental manuscripts on the ethical dimensions of AI identified by our adapted discourse approach. We interpret these findings and discuss the role of IS research in the ecosystem of the ethical dimensions of AI in Section 5. We provide concrete contributions to IS research and highlight an avenue for future studies. We conclude with closing thoughts and a call to action in Section 6, reflecting on how scholars may build upon our results.

2 Research Background

Ethics scholarship in IS deals with various questions, such as privacy, intellectual property, employment relationships, design decisions, and the changing role of humans in society (Stahl, 2008). As early as 1985, Moor (1985) distinguished computer ethics from ethics in relation to other technologies. In this context, most research on ethics deals with normative challenges (Stahl, 2008). Thus, illegal, inappropriate, and unethical behavior is researched in the context of information technology (Leonard et al., 2001; Sojer et al., 2014). Recommendations for action, agendas, or frameworks for ethical research and practice are therefore established (Stahl, 2008; Stahl et al., 2014; Walsham, 1996).

Computer and algorithm biases are a curated ethical issue in IS research. There are several types of biases in AI technologies, such as sampling bias, which produces models relying on training data that is not representative of future cases, and performance bias, which examines performance distortion in predictions by AI (Abbasi et al., 2018). In addition, confirmation bias can lead to machine learning searches that reinforce biases, and anchoring bias can lead to incorrect assumptions about initial information provided by AI. An established classification of computer bias is a framework provided by Friedman and Nissenbaum (1996). They defined criteria such as reliability, accuracy, and efficiency by which the ethical quality of computer systems should be judged. Ethical principles were also established for specific methods of IS research. One example can be found in the ethical principles for design science research, which are: public interest, informed consent, privacy, honesty and accuracy, property, and quality of the artifact (Myers & Venable, 2014). It is important to discuss these principles because violating them can cause harm to individuals or society. Furthermore, compliance with these principles can improve social coexistence or reduce discrimination against individuals. The same principles can also be found in other contexts, such as ethical guidelines for internet communities (King, 1996). More recent research on information privacy in organizations has also considered the constructs of control, justice, and ethical obligation (Greenaway et al., 2015). These principles were transferred to concepts such as nudging (the guiding of individuals' behavior toward a beneficial choice for themselves or society) and have been expanded accordingly (Renaud & Zimmermann, 2018). The ethical principles for nudging are 1) respect (including retention and transparency), 2) beneficence, 3) justice, 4) scientific integrity, and 5) social responsibility. In the current discourse on the ethical dimensions of AI, these principles have again been used and extended to transfer them to autonomous computer systems (Floridi & Cowls, 2019). The discussion of these principles in the context of AI technologies is highly unstructured, especially in IS research, and has so far only scratched the surface of an important social challenge. There are no fundamental IS works that provide directions for future research on the ethical dimensions of AI. The conflict between ethics and AI has not been addressed sufficiently in IS research, leaving ethical issues unresolved which could impact people's lives. Problems such as privacy abuses or hate speech that have arisen in connection with social media technologies show that it is important to create ethical frameworks prior to the widespread utilization of new technology. As AI will penetrate more and more areas of professional, public, and private life in the future, it is important to prevent possible damage to society and individuals, to maximize its benefits, and to guide developments ethically. IS scholars can take a leading role in this quest due to their expertise in understanding socio-technical phenomena.

IS research has a long history of examining and ensuring the ethical use of computers and curating this knowledge (Chatterjee et al., 2009; Kallman, 1992; Stahl, 2008). Various frameworks, principles, and guidelines have been established to support researchers and practitioners in the ethical use of computers (Ess, 2009; Harrington, 1996; King, 1996; Sojer et al., 2014). Nonetheless, Stahl noted in 2008 that there were only a small number of IS papers dealing with ethics. Although the research interest in the ethics aspects of IS grows continuously, there remains a dearth of fundamental articles on emerging technologies such as AI-based technologies.

2.1 Ethics and AI

Currently, research on the ethical dimensions of AI is trending. AI ethics differs from the debate on other technologies, as AI raises different ethical questions in relation to other technological trends, such as blockchain, big data, or virtual reality (Boddington, 2017). As summarized by Russel and Norvig (2016), AI can be defined as a research stream that includes all technologies that can think or act like a human or that can think or act rationally. However, not only do the capabilities of AI-based systems continue to evolve, but so does what can be defined as AI. Currently, AI can be considered a frontier of computational advancements, capable of solving more and more complex decision problems that were once reserved for humans (Berente et al., 2021). In practice, and in most IS case studies, AI is usually considered to be a technology with self-learning abilities via machine learning, neuronal networks, or deep learning and thus performs better than a human in narrow tasks (Brynjolfsson & Mitchell, 2017; Kotsiantis, 2007). AI can thereby relieve people from repetitive processes (Dias et al., 2019). However, unlike most other technologies, AI not only threatens the jobs of employees who perform many repetitive tasks but can also replace the work of knowledge workers by being designed to make independent decisions (Boddington, 2017). Furthermore, studies show that many people already perceive AI as an independent individual (Araujo, 2018; Feine et al., 2019; Mirbabaie et al., 2021b; Seeber et al., 2020), which also raises ethical questions. In addition, the use of AI is not an exact science, since AI learns and builds on predictions (Boddington, 2017). AI technologies are usually trained on huge datasets that can hardly be traced by a human. This means that the output of AI cannot always be easily explained. Due to the complex algorithms and the huge amount of training and test data, AI takes on the form of a certain "black box" like character for humans, resulting in difficulties tracing back the outputs of AI predictions. In particular, when AI has to make important decisions that impact directly on a person’s life (such as getting credit approval or health insurance), major ethical challenges arise (Aversa et al., 2018; McNamara et al., 2018). The research on Explainable AI addresses this topic (Barredo Arrieta et al., 2020; Gunning, 2017; Miller et al., 2017). Furthermore, ethical dilemmas arise when an AI makes a challenging decision and a human is directly or indirectly harmed by it (Coppersmith, 2019). For example, in the case of an accident involving a self-driving car, the question arises as to whether responsibility for the damage lies with the developer, the supplier, the customer, or even the technology itself.

In order to address these socially relevant ethical problems with AI, governments and organizations have established guidelines and policies on how AI should be used. One example can be found in the Ethics Guidelines for Trustworthy AI, which were formed by a European Union expert committee to regulate the use of AI in European countries (EU HLEG, 2019). Other countries, such as China or Canada, also have their own guidelines for the use of AI (BAAI, 2019; Floridi et al., 2018). In addition, large organizations such as Google, Microsoft, and IBM have published guidelines (Vakkuri et al., 2019).

Research on this topic is in its early stages, needing guidance and a clear understanding of the cumulative tradition of related domains and what needs to be addressed in future research. An initial attempt to synthesize the various principles was carried out by Floridi et al. (2018). The authors identified four core risks of AI: devaluing human skills, removing human responsibilities, reducing human control, and eroding human self-determination. Furthermore, they established a framework and recommendations for a good AI society, considering the AI guidelines of various governments (e.g., The Montreal Declaration for Responsible AI) and institutions (e.g., Asilomar AI Principles) (Floridi et al., 2018). As core principles for the ethical use of AI, the same researchers identified beneficence, non-maleficence, autonomy, justice, and explicability.

2.2 Discourse on the Ethical Dimensions of AI in IS Research

Research on the ethical dimensions of AI is a broad field, including a wide range of disciplines. de Almeida et al. (2020) provided an overview of frameworks and guidelines on the ethical dimensions of AI. They carried out a systematic literature review including peer-reviewed articles from several relevant databases using keywords such as “ethics”, “how to regulate”, “risk”, and “framework”. Although they offered a broad overview of frameworks on AI ethics in IS research and beyond, their review was limited to peer-reviewed articles published in journals in a given time period using specific keywords for their identification (de Almeida et al., 2020).

However, even within IS research, empirical and theoretical works hardly differ in terms of their viewpoints on the ethical dimensions of AI. Thus, the implementation of values such as power, achievement, hedonism, stimulation, self-direction, universalism, benevolence, tradition, conformity, and security were examined (Robbins & Wallace, 2007). In addition, the problem of bias in machine learning is a trending focus in IS research (Ahsen et al., 2019; Kordzadeh & Ghasemaghaei, 2021; Yapo & Weiss, 2018) and in management practice (Martin, 2019). Other IS works on ethical dimensions of AI focus exclusively on specific application domains such as hiring (Hofeditz et al., 2022) or healthcare (Mirbabaie et al., 2021a).

Furthermore, Teodorescu et al. (2021) highlighted failures of applying fairness in human-AI augmentation, resulting in unintentional discrimination. They argued that IS scholars’ knowledge on how to address the principle of fairness for AI-based systems is limited and call for further research. On another level, Etzioni and Etzioni (2016) suggested that a new variety of AI technologies should ensure that existing AI-based systems meet ethical standards by monitoring, auditing, and holding operational AI systems accountable. Porra et al. (2019) argued that it will most likely turn out not to be beneficial for our societies if AI becomes increasingly anthropogenic. They predicted that digital assistants would outnumber humans by 2021, and, therefore, the ethical dimensions of AI should be discussed philosophically (Porra et al., 2019). In 2021, the market for digital assistants continues to grow strongly (Research and Markets, 2021).

In sum, the discourse on the ethical dimensions of AI in total, and especially in IS, is taking place on different levels of abstraction and from various empirical and theoretical angles. Previous reviews reflect parts of the big picture. However, it is unclear which manuscripts are fundamental for the research domain and which future research directions scholars should address.

3 Research Design: An Adapted Discourse Approach

To the best of our knowledge, the ethical dimensions of AI are ambiguous and the discourse on how to address the ethical issues of AI is unstructured. A systematic literature review is a method to reveal the current state of the art on a theory and to point out gaps or define a research agenda (vom Brocke et al., 2015). However, according to Larsen et al. (2019), it is hardly possible to identify and properly consider all relevant works due to the constant growth in knowledge and the sometimes very high number of publications on a topic or a theory. Therefore, we use a discourse approach, which starts from fundamental theory-building papers (L1) that derived a fundamental theory, framework, or model or that shed light on a phenomenon or a new research domain. As a second step, theory-contributing papers and papers that cited the L1 articles will be identified (L2). The third type of papers (L3) are those that influenced the L2 papers. L1, L2, and L3 form the interconnected ecosystem of a theory or domain.

We have adopted this approach for our review of the research field of the ethical dimensions of AI in order to structure the discourse and understand the ecosystem behind it. Larsen et al. (2019) did not describe in their work how they identified fundamental manuscripts (L1 papers). In our case, the fundamental manuscripts were not apparent at first. Therefore, we developed a method to be able to identify L1 papers. Our research approach consists of three phases, following the recommendations of Larsen et al. (2019). An overview of the applied research approach is provided in Figure 1 and will be presented in the following sub-sections.

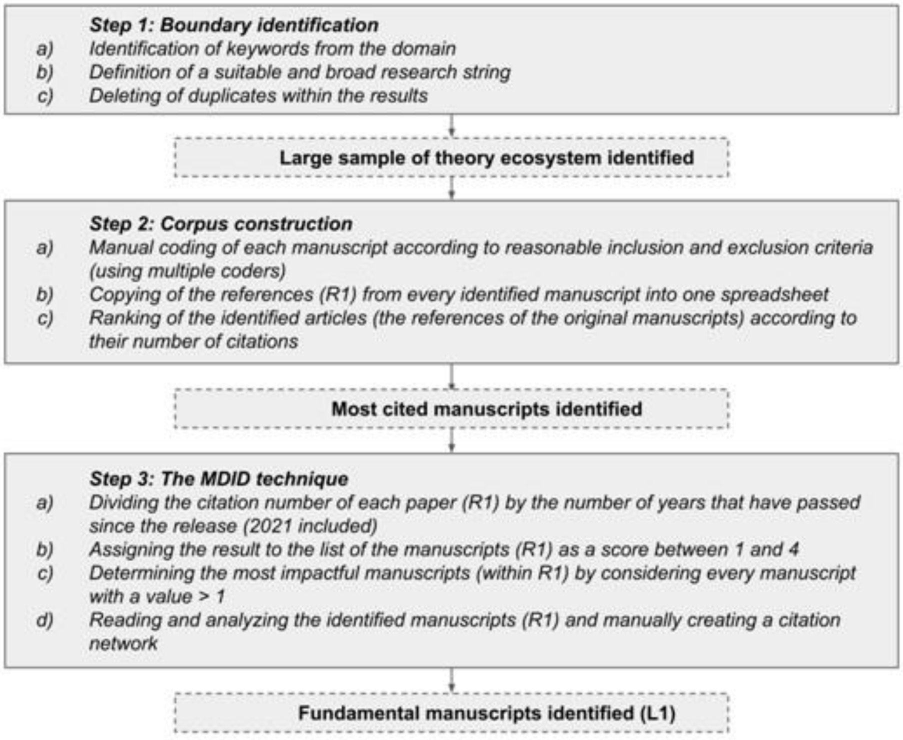

Step 1: Boundary identification
a) Identification of keywords from the domain
b) Definition of a suitable and broad research string
c) Deleting of duplicates within the results
Large sample of theory ecosystem identified
Step 2: Corpus construction
a) Manual coding of each manuscript according to reasonable inclusion and exclusion criteria (using multiple coders)
b) Copying of the references (R1) from every identified manuscript into one spreadsheet
c) Ranking of the identified articles (the references of the original manuscripts) according to their number of citations
Most cited manuscripts identified
Step 3: The MDID technique
a) Dividing the citation number of each paper (R1) by the number of years that have passed since the release (2021 included)
b) Assigning the result to the list of the manuscripts (R1) as a score between 1 and 4
c) Determining the most impactful manuscripts (within R1) by considering every manuscript with a value > 1
d) Reading and analyzing the identified manuscripts (R1) and manually creating a citation network
Fundamental manuscripts identified (L1)

Figure 1. Adapted Discourse Approach In Three Steps (Source: Larsen et al., 2019).

3.1 Boundary Identification

The first step in every literature review should be the identification of boundaries. Systematic literature reviews have been regularly criticized for not providing a comprehensive picture of a discourse (Larsen et al., 2019; vom Brocke et al., 2015). The discourse approach of Larsen et al. (2019), however, considers a research domain less as a set of characteristics but rather as a discourse between scholars. As the aim of our work was to gather a literature base and to identify the origin of the current discourse on the ethical dimensions of AI, we applied the approach in order to derive directions for IS and IS-related research. To do this, however, we did not limit our search to IS journals and conference papers alone, but also to relevant manuscripts from outside IS research. The approach is centered on so-called L1 papers, which represent the best, most cited, or most well-known papers in their respective research stream. These L1 papers are fundamental manuscripts about a theory, a model, a framework, a research domain, or a (trending) topic. For instance, Davis (1989) was mentioned as an example of a fundamental L1 paper on the Technology Acceptance Model (TAM).

Papers that want to contribute to academic discourse and develop a theory or domain should cite fundamental manuscripts (Larsen et al., 2019). From the citations of the theory-contributing papers and the papers on which these manuscripts are based, a citation network can be developed, which can be called a theory or domain ecosystem.

To define this network, the L1 papers must be identified first. However, Larsen et al. (2019) do not describe an exact process for tracking fundamental papers. In their example, the L1 paper was presented as generally known (the TAM by Davis (1986)). For new research domains and emerging fields and phenomena, however, there often is no consensus on the origin of a discourse. Thus, in the context of the ethical dimensions of AI, a predefined set of fundamental papers (L1) has yet to be identified in order to identify contributing articles (L2). Therefore, we had to modify the discourse approach in order to identify papers that can be considered fundamental for the discourse of the ethical dimensions of AI. Hence, we decided to commence by applying a "traditional" systematic keyword search but with a broad search query. We used as many synonyms as possible for terms from our domain of interest in order to follow Larsen et al. (2019), who recommended not limiting the search to a too narrow search string. However, our aim was to identify the status quo in IS research regarding the ethical dimensions of AI; therefore, our starting point for our search was based in IS research. The following search string was run through Scopus (with the help of Litbaskets.io[1] to identify relevant IS journals and IS-related interdisciplinary journals) and AISeL databases (mainly to include IS conference pieces):

("artificial intelligence" OR "AI") AND ("ethics" OR "ethical" OR "ethic")

We know that there are synonyms for both artificial intelligence (such as "machine learning" or "neural networks") and ethics (such as "morals" or "morality"), but through an upstream keyword search, we found that all articles that really discussed ethical dimensions of AI and not only one facet (such as ethics in IS or AI) contained the keywords of "artificial intelligence" and "ethics" or "ethical". We started our initial search by choosing AISeL (mainly for IS conferences) and Scopus as a meta database (for the Basket-of-Eight journals, general IS journals, and IS-related journals) as we wanted to contribute to the ongoing discourse on the ethical dimensions of AI in IS, and these databases include all manuscripts such as journal articles, conference proceedings, and books that can be considered IS research. However, only the starting point of our search was focused on the IS discipline to identify the status quo in IS research. The further steps of the systematic search, including a forward and backward search according to Webster and Watson (2002), identified articles published outside IS. However, we deem those papers relevant as they are related to the discourse on the ethical dimensions of AI.

We performed our literature search between June and July 2020, and it was updated during the revision process. According to Larsen et al. (2019), we did not apply any filter or limitations by year. As a next step, we identified duplicates within the results. Our initial search resulted in 381 papers. After deleting duplicates, we ended up with 175 results. As none of these articles stand out by the number of citations per year, a holistic view, or the connectedness within the results, we assumed that these articles can be labeled as L2 or even L3 articles. With these results, we were still not able to understand the discourse on the ethical dimensions of AI or even determine the center of the discourse by identifying L1 manuscripts. Although the keyword search was a necessary first step to shed light on the discourse on the ethical

[1] Litbaskets is an information technology artifact supporting exploratory literature searches for information systems research (Boell & Wang, 2019).

dimensions of AI, we also assumed that our initial keyword search did not precisely cover the whole picture for a corpus of literature. Therefore, we proceeded with a more comprehensive cross-disciplinary search to understand the discourse on the ethical dimensions of AI and identified the fundamental (L1) manuscripts.

3.2 Corpus Construction

To build a corpus of literature, the manuscripts of the initial search must be considered in more detail. As is the case for all literature reviews, not all manuscripts are relevant (Larsen et al., 2019). For the next step, two independent coders filtered the papers for relevance and fit to our topic, applying inclusion and exclusion criteria. The coders manually scanned abstracts and keywords from the population identified. We included articles that added new knowledge to the discourse on AI and ethics or contributed to existing guidelines, models, or frameworks. We excluded articles that mentioned ethical aspects or AI just as a side note. To measure the intercoder reliability, we used Cohen's κ. We calculated an intercoder reliability of κ = 0.91. This step led to 125 relevant papers from the initial search. Since none of the papers still stood out, and we did not find a close connection between those articles, but because they all addressed the ethical dimensions of AI, we classified these papers as L2 (discourse contributing).

As with most research, these articles contributed to a research domain by citing and discussing certain previous works. We concluded that if all articles contribute to the discourse on ethical dimensions of AI but none of the articles within the corpus could be considered as fundamental manuscripts, fundamental works need to be among the references of those articles. Therefore, we copied all references of these 125 papers from the identified population into a list, which led to a total of 5,077 references that were no longer limited to IS research and contained manuscripts of various disciplines. L1 papers are manuscripts that should be cited in many articles addressing the discourse on a research domain. Therefore, we ranked the identified manuscripts in the reference table according to how often they were cited by the initially identified papers (which was not equal to the total number of citations, e.g., on Google Scholar or Scopus). After checking these references manually, we came up with the results presented in Table 1. These papers can be considered highly relevant for our research domain, although most of them cannot be allocated to the IS discipline. However, our aim was to understand the current interdisciplinary discourse on the ethical dimensions of AI to derive directions for IS research, and we neither knew a threshold for which articles need to be discussed in more detail nor did we consider the year of publication in relation to the number of citations within the 125 identified articles. Inspired by Larsen et al. (2019), we, therefore, developed a manual detection of implicit domain (MDID) technique to identify articles that came closest to what Larsen and his colleagues described as fundamental manuscripts for discourse in research.

Table 1. Ranking of Identified Articles According to their Number of Citations[2]

Number of citations within the 125 identified articles	Number of papers
14	1
12	2
8	4
7	5
6	1
5	6
4	19
3	64

[2] Papers that were quoted less than five times throughout the 5,077 references were omitted from this initial count due to time constraints.

Table 1. Ranking of Identified Articles According to their Number of Citations[2]

2	226
1	2105

3.3 A Manual Detection of Implicit Domain (MDID) Technique

To identify theory-contributing manuscripts in an ecosystem, Larsen et al. (2019) used an automated detection of implicit theory technique based on machine learning. However, we were not able to detect at least one L1 paper for the discourse on the ethical dimensions of AI accurately. In this article, we, therefore, developed a manual technique based on a ranking of citations among articles identified by a systematic keyword search to then be able to identify fundamental manuscripts and highlight the literature ecosystem (Larsen et al., 2019) of the ethical dimensions of AI across disciplines to guide IS research.

After reviewing the most cited manuscripts within the references of our initially interdisciplinary searched papers (not the total number of citations), we found that there was a wide time span between the publication dates of the manuscripts. However, we aimed to understand the current discourse on the ethical dimensions of AI, as for such an emerging field, a discourse can change its focus over time. Therefore, we divided the number of times articles occurred within the reference lists of the initially identified 125 articles by the number of years that have passed since the release of their first version, 2020 included. This led to a value between 0 and 4 citations per year, with only a few papers in the range of 1 to 4 and many papers at 1 or below. We considered these values as a score that describes the impact of the manuscripts on the current discourse on the ethical dimensions of AI. To be able to determine a threshold for the most relevant articles, we visualized the number of times these articles were cited by the 125 initial articles on a graph. An excerpt of this graph is provided in Figure 2, which shows the distribution of the scores of the manuscripts and some examples of paper titles.

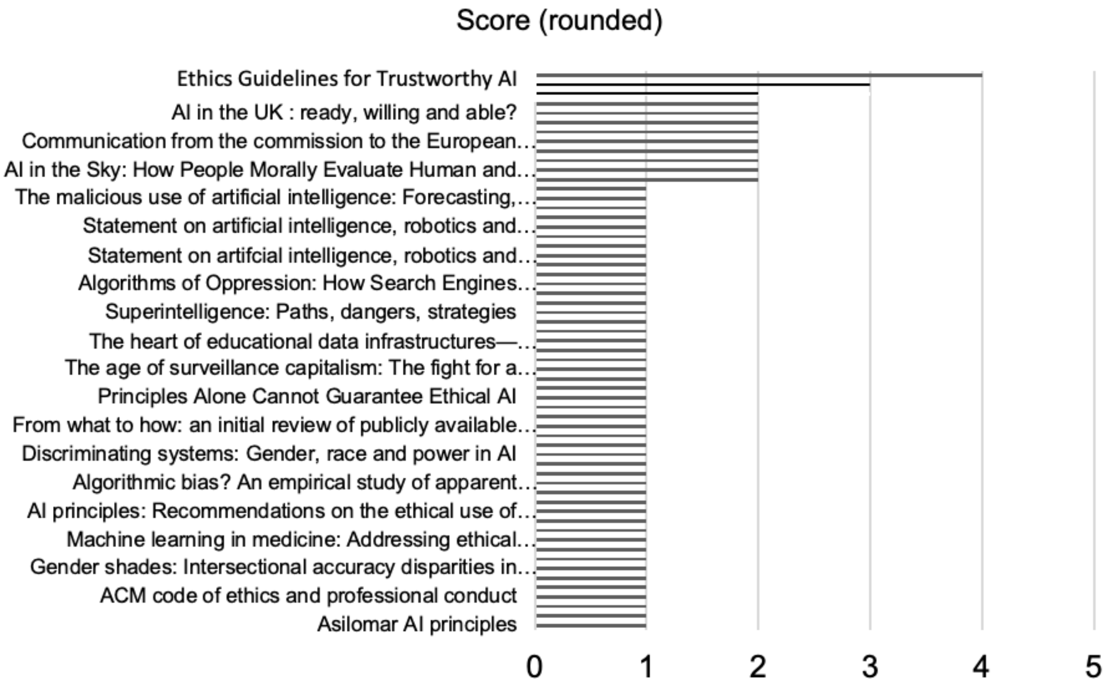


Figure 2. Visualization of an Extract from the Distribution of the Scores of the Identified Papers.

After reviewing the data and visualizing the distribution of referred manuscripts on a graph, we assessed every paper with a score above 1 to be impactful enough to be called a fundamental manuscript (as they visually stood out on the graph), which led to a total of 12 papers. Table 2 provides an overview of the manuscripts that came closest to what Larsen et al. (2019) described as fundamental L1 manuscripts. We described which artifacts were discussed in these manuscripts and compared the calculated scores with the overall citations on Google Scholar.

It was not our aim to extract the complete ecosystem by classifying every single paper in a citation network. We focused on the origin of the current discourse on the ethical dimensions of AI. However, within this process, we also identified some contributing L2 manuscripts.

Table 2. Literature Classified as L1 by Applying MDID Technique, Sorted by Score. (Status: February 2022)

ID	Consideration/Artifact	Author & Year	Outlet	Score (rounded)	Cit. in sample	Schol. Cit.
#01	Ethics Guidelines for Trustworthy AI	(EU HLEG, 2019)	EC Europe	4	8	0
#02	Ethical Framework for a Good AI Society	(Floridi et al., 2018)	Minds and Machines	3	8	679
#03	“Weapons” of Math Destruction	(O’Neil, 2016)	Broadway Books	2	12	4411
#04	AI recommendations for the UK	(House of Lords, 2018)	House of Lords (UK parliament)	2	7	0
#05	ACM’s Code of Ethics	(McNamara et al., 2018)	ESEC/FSE 2018 (conference)	2	5	105
#06	Metareview on researching algorithms	(Kitchin, 2017)	Information, Communication & Society	2	6	836
#07	Case studies for AI in military	(Malle et al., 2019)	Robotics and Well-Being	2	3	35
#08	Recommendations for AI in healthcare	(Yu et al., 2018)	Nature Biomedical Engineering	2	3	670
#09	Beijing AI principles	(BAAI, 2019)	BAAI	2	3	0
#10	Industry viewpoint and an empirical study on ethically aligned design of autonomous systems	(Vakkuri et al., 2019)	Computers & Society	2	3	22
#11	Overview of AI ethics tools, methods and research to translate principles into practices	(Morley et al., 2020)	Science and Engineering Ethics	2	3	199
#12	Ethically Aligned Design (EAD v1 & v2)	(Shahriari & Shahriari, 2017)	2017 IEEE Canada International Humanitarian Technology Conference (IHTC)	2	8	34

Furthermore, supplementing the numeric analysis, we manually checked each paper for its relevance and its role in the domain ecosystem. We extracted frameworks, guidelines, models, theories, and theory-contributing work in order to illuminate the discourse on AI ethics. In addition, we visualized how our fundamental manuscripts were cited and cited each other.

4 Results

Although we commenced our discourse approach from an IS perspective using IS databases (before we expanded our search to other disciplines through both forward and backward search), we did not find one fundamental paper published in an IS journal or in IS conference proceedings among the most-cited articles in our cross-disciplinary systematic search. Many of the most frequently mentioned manuscripts among the papers of our identified corpus were reports, books, or white papers from governmental or research institutions. With our interdisciplinary MDID technique, we also found research papers from other disciplines that could be considered fundamental for the discourse on the ethical dimensions of AI.

Before we applied our weighting of the citations, one article stood out, as it was cited 14 times by the papers that we identified with our keyword search. Turing's seminal paper on AI addressed the question of whether AI can or will ever be able to think like humans (Turing, 1950). Within his work, he introduced the "imitation game," also known as the Turing Test. Although the paper was the first seminal work on the ethical dimensions of AI, we did not consider it a fundamental L1 manuscript for the current discourse due to the score we used to weight the identified papers. Below, we discuss those manuscripts that we classified as L1 papers after applying our MDID technique.

One of the most frequently cited manuscripts we identified in our domain ecosystem was the EAD guidelines (v1 & v2) published by a committee of the IEEE Global Initiative (Shahriari & Shahriari, 2017). The document, of which there now exists an updated version, was developed based on the knowledge of several hundred leaders from six continents from academia, industry, civil society, politics, and government. Their aim was to enable ethical and social implementations of AI technologies in accordance with human values and ethical principles. Furthermore, the guidelines were intended to encourage researchers to develop new standards. Fundamental principles include the embodiment of the highest ideals of human beneficence as a superset of human rights and the prioritization of people and the natural environment when applying AI. In addition, risks and negative influences, as well as misuse, should be mitigated through transparency and accountability. As these IEEE guidelines were one of the two most prominent artifacts within our ecosystem, we classified the manuscript as L1.

The "Ethics Guidelines for Trustworthy AI" were quoted very frequently and achieved the highest score overall. The guidelines were established by the EU Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) as part of the European AI strategy (EU HLEG, 2019). The manuscript contains the Framework for Trustworthy AI, which we classify as one of the fundamental L1 frameworks for the considered research domain of AI ethics. The framework is based on four basic principles: 1) respect for human autonomy, 2) prevention of harm, 3) fairness, and 4) explicability. In addition, eight key requirements should be fulfilled before an AI can be implemented: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing, and accountability. These requirements are of high importance in the field of AI, as AI technologies tend to have more autonomy in decision making and can, therefore, cause greater harm to humans than most other technologies. AI is constantly evolving, and its outputs are hardly traceable for humans, which can result in errors being detected very late.

As another fundamental manuscript on the ethical dimensions of AI, we identified the article by Floridi et al. (2018) that we already highlighted in the research background. The manuscript reports the results of the AI4People initiative, which aims to create a foundation for a good AI society. The researchers identified beneficence, non-maleficence, autonomy, justice, and explicability as basic principles for the ethical use of AI. They also formulated 20 concrete recommendations for the development, incentives, and support of good AI. The paper lists more than 200 Google Scholar citations and showed a very high relevance within the domain ecosystem (Floridi et al., 2018). Therefore, we also classified it as a fundamental L1 paper.

In "Weapons of Math Destruction," O'Neil (2017) argues that decisions affecting people's lives will increasingly be made using mathematical models (Verma, 2019). This results in less fairness, as these models are opaque, unregulated, and incontestable. The book was difficult to categorize in the domain's ecosystem, as it primarily addresses Big Data rather than AI. However, since the book has been cited frequently as a basis for further IS research and achieved a high score, we classified it as an L1 work. The two manuscripts #05 and #10 were reports and recommendations of the British and Chinese governments, respectively, on the use of AI. In #05, the recommendations of the British AI Council, the Centre for Data Ethics and Innovations, the Alan Turing Institute, and the Government Office for AI were merged into one document of guidelines (House of Lords, 2018). The recommendations for action in #10 were divided into three areas: 1) research and development, 2) use, and 3) governance. These principles were developed by the Beijing Academy of Artificial Intelligence (BAAI) and are being used by leading research institutions and organizations in China (BAAI, 2019). Therefore, we classified both #5 and #10 as L1 manuscripts.

We also found the ACM's code of ethics to be a fundamental framework (McNamara et al., 2018). We classified the code and the conference paper identified in our search as L1 manuscripts, as it achieved a score of 2. The ACM's code of ethics primarily aims at guiding researchers and practitioners in the field of computer science. The principles are divided into three sections: 1) general principles, 2) professional leadership principles, and 3) compliance with the code. They are formulated very broadly and include, for example, the following phrase: “Be fair and take action not to discriminate.” Although a study has implied that consideration of the ACM’s code of ethics has no effect on decision making, it is a fundamental manuscript for the domain ecosystem (McNamara et al., 2018).

Manuscripts #07, #08, and #09 did not consider fundamental theories, frameworks, or models of ethical AI, nor did they contribute to any of the fundamental manuscripts already identified. Rather, they discussed subdomains such as AI ethics in healthcare (Yu et al., 2019) and narrow challenges for ethical AI such as AI's ethical dilemmas in military operations (Malle et al., 2019). Nevertheless, important ethical challenges and issues regarding the use of AI were addressed, and the manuscripts achieved a high score according to the weighted citations. We found many articles (L2) that build on the findings of these manuscripts. These manuscripts address key areas that are not covered in the other L1 manuscripts. Just among the other fundamental manuscripts, they were not discussed. As they can be considered pioneering work on the ethical dimensions of AI, we classified these papers as manuscripts that came close to L1 manuscripts.

Vakkuri et al. (2019) conducted a large empirical study. They conducted a multiple case study with five organizations to demonstrate a gap between research and practice on AI ethics, further providing recommendations for closing this gap. They referred to ACM’s code of ethics, the Ethics Guidelines for Trustworthy AI, and the guidelines on Ethically Aligned Design. But even within the population of the initially identified articles, the manuscript was cited frequently and reached a score of 2. Therefore, we classified the manuscript as a theory- and domain-contributing L1 paper.

In the last manuscript we identified as fundamental, Morley et al. (2020) argue that the discourse on AI ethics focuses too much on principles and too little on practices. They also attempt to close the gap between principles and practices, referring to the conclusions and recommendations of the British House of Lords (House of Lords, 2018), the Ethical Framework for AI (Floridi et al., 2018), the Ethical Guidelines for Trustworthy AI (EU HLEG, 2019), and the framework of Ethically Aligned Design (Shahriari & Shahriari, 2017). They also refer to further guidelines and principles such as Asilomar’s AI Principles and IBM’s Everyday Ethics for AI (Morley et al., 2020). In addition, their work has been cited frequently and reached a score of 2. Therefore, we also classified this manuscript as theory contributing L1 work.

In total, we were able to classify all 12 manuscripts we identified by our adapted discourse approach as fundamental L1 papers since they either provide guidelines, principles, or frameworks on the ethical dimensions of AI or address them. However, only four of the identified fundamental manuscripts were peer-reviewed journal articles, and two were conference proceedings. Furthermore, no article published within the IS community could be recognized. Seven of the articles did not establish new frameworks but rather discussed existing guidelines and frameworks or narrow subdomains. Except for five of the papers, the manuscripts referred to at least one other L1 article or report. These five manuscripts did not refer to other fundamental papers but discussed AI ethics either on a meta level or addressed practical challenges or AI dilemmas. The 12 identified L1 manuscripts are visualized as a chronologically sorted citation network in Figure 1. The arrows indicate how the manuscripts cited each other. The citations mentioned in the figure are the Google Scholar citations from August 2020.

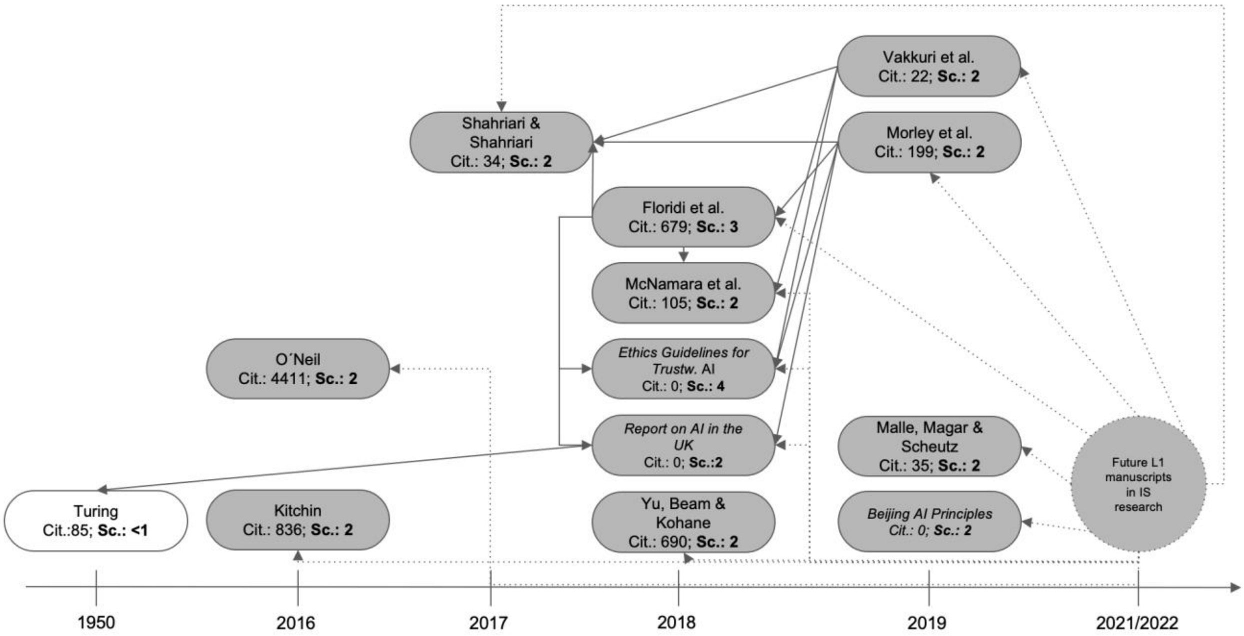


Figure 3. Domain Ecosystem for the Current Discourse on the Ethical Dimensions of AI.

The gray boxes in Figure 3 represent the manuscripts extracted from our identified corpus. The manuscripts also discuss other AI principles, such as Google's AI principles, IBM's Everyday Ethics for AI, Microsoft's guidelines for conversational bots, Intel's recommendations for public policy principles on AI, the Montreal Declaration for Responsible AI, and the Future of Life's Asilomar AI principles (Morley et al., 2020). In addition, Turing's article on the imitation game was cited the most among the considered manuscripts. However, it did not achieve a high enough score to be classified as L1, which is why we visualized the paper in a white box. The circle on the bottom right in the figure highlights possible future fundamental IS papers on the ethical dimensions of AI.

5 Discussion

Literature reviews are essential to structure an ongoing discourse or to provide research directions. Nevertheless, the method of the literature review needs to be developed further (Larsen et al., 2019; Rzepka & Berger, 2018; vom Brocke et al., 2015). The discourse approach of Larsen et al. (2019) is one of the latest methods to structure a discourse on a theory using reverse citations. In this approach, a network of citations is built from fundamental L1 manuscripts. However, as described by the authors, there is not always such a clearly defined point of origin. The discourse on the ethical dimensions of AI is such a discourse without a clear origin. Larsen et al. (2019) did not provide information on how the approach can be applied in such a case. However, since the discourse approach is based on citations, we followed this argument and offered a solution to identify fundamental manuscripts when they are initially unclear.

5.1 Discussing the Ethical Dimensions of AI

Our adapted discourse approach was well suited to identifying fundamental manuscripts on the ethical dimensions of AI. Overall, the MDID technique worked quite well to identify the most important manuscripts in the domain ecosystem. Interestingly, the papers we identified were quite different from the most cited papers on AI ethics in a simple Google Scholar keyword search. Some of the articles found by Google Scholar may also be important manuscripts; however, they rarely or never appear in the core ecosystem of the ethical dimensions of AI. It should also be noted that Google Scholar, as well as other literature databases such as Scopus, do not contain all important manuscripts for a comprehensive theory or domain ecosystem. That is why the relevance of articles within a research domain cannot be determined by citations in a database. Furthermore, often applied exclusion criteria in keyword searches, such as limitations by certain years, (specific) journal articles, or peer-reviewed articles only, lead to an incomplete picture of a discourse. Within our corpus, documents such as the Beijing AI principles and the report on AI in the UK were highly relevant, despite not being listed in the common literature databases. Thus, we agree with Larsen et al. (2019) that it is important that no manuscripts are excluded from the initial literature search. However, the score we developed not only enabled us to illustrate the discourse on a research domain and to identify L1 articles, but we could also identify the most relevant manuscripts for the current discourse.

Although we initially started our literature search in the corpus of IS, we did not find any IS journals or conference proceedings among the manuscripts we identified as fundamental. The only IS-related research article we could classify as L1 was published in a philosophy journal (Floridi et al., 2018). Overall, no single discipline can be identified that forms the origin of the current discourse on the ethical dimensions of AI. Nevertheless, we have found that most of the fundamental manuscripts originate from the disciplines of philosophy and computer science. Although one of the most important fundamental works is still Alan Turing’s work on the imitation game (Turing, 1950), a new generation of fundamental manuscripts is now emerging in the domain.

We found that many of the manuscripts we classified as L1 were reports and recommendations from governments, institutions, or organizations. These contained guidelines, frameworks, principles, or recommendations for action. According to Larsen et al. (2019), we included conference proceedings and preprints in our corpus, which proved to be very valuable. We identified two fundamental manuscripts that were conference proceedings and one preprint published on arXiv that would most likely be excluded in a traditional systematic literature review process such as the one described by vom Brocke et al. (2015).

Although the identified manuscripts from our domain ecosystem refer to each other, there is no superordinate L1 paper covering the entire spectrum of the domain. The most relevant manuscripts among the 12 fundamental papers were those of Floridi et al. (2018), the Ethically Aligned Design (EAD v1 and v2), the Ethical Framework for AI, and the ACM’s code of ethics. These documents have many similarities. The principles of explainability, prevention of harm, and respect for human rights are used as basic principles in most guidelines. In addition, benefits, autonomy, and justice are often mentioned, referring to the traditional principles of bioethics (Floridi et al., 2018). Some frameworks also refer to the practical readiness for AI ethics of organizations (Floridi et al., 2019). Interestingly, the AI ethics principles of the Chinese government are also strongly aligned with the values of Western cultures.

Although IS literature was not found among the fundamental manuscripts for the ethical discourse on AI, it indirectly contributed to its development. Except for the principle of non-beneficence, we found a similar counterpart for each ethical dimension of AI within the IS literature. Non-beneficence or prevention of harm also appears in a more moderate IS beneficence principle (Renaud & Zimmermann, 2018) or is described as a general “control” principle (Myers & Venable, 2014). However, the principle is particularly relevant to AI, as AI technologies are now and will continue to be given significantly more decision-making power than other technologies have ever had in the past (Floridi et al., 2018). We transferred the IS ethics principles and the ethical principles of AI into the ethical dimensions of AI that aim to guide future research and development of AI. The dimensions are visualized in Table 3.

Table 3. Comparison of IS Ethics Principles and Ethics Principles for AI.

Traditional IS ethics principles	Ethics principles for AI	Ethical dimensions of AI
Beneficence Renaud & Zimmermann (2018)	**Beneficence** Floridi et al. (2018), McNamara et al. (2018), Shahriari & Shahriari (2017) **For Humanity** (BAAI 2019)	Researching and developing AI should **contribute to the common good** and should consider privacy, dignity, freedom, autonomy, and rights of users.
Beneficence Renaud & Zimmermann (2018)	**Non-Maleficence** EU HLEG (2019), Floridi et al. (2018) **Prevent Harm to Humans** BAAI (2019), McNamara et al. (2018)	When researching and developing AI, **misuse should be prevented,** and caution should be implemented to avoid harm to humans.
Justice/Transparency/Respect Greenaway et al. (2015), Renaud & Zimmermann (2018)	**Justice/Explicability** EU HLEG (2019), Floridi et al. (2018), House of Lords (2018), McNamara et al. (2018), Shahriari &	Research and development of AI should be as fair as possible and **reduce possible discrimination**. Transparency and explainability

Table 3. Comparison of IS Ethics Principles and Ethics Principles for AI.

	Shahriari (2017) **Debiasing** BAAI (2019)	should be as high as possible in order to **prevent biases**. Make AI more explainable, predictable, traceable, auditable, and accountable.
Public Interests King (1996), Myers & Venable (2014)	**Do Good** BAAI (2019), EU HLEG (2019), Floridi et al. (2018), McNamara et al. (2018)	Researchers and developers of AI should enhance the well-being of society and ecology. Therefore, **stakeholders who may be affected need to be identified**. Security, autonomy, health, democracy, empowerment, and anticipation should be placed above features and capabilities.
Control Greenaway et al. (2015), Myers & Venable (2014)	**Autonomy** Floridi et al. (2018), EU HLEG (2019), Shahriari & Shahriari (2017), EU HLEG (2019), Floridi et al. (2018), Shahriari & Shahriari (2017)	Researchers and developers should **ensure that users have a certain level of control** when interacting with an AI.
Quality of the Artifact Greenaway et al. (2015), Myers & Venable (2014)	**Control Risks** BAAI (2019), House of Lords (2018)	Researchers and developers should **improve the maturity, robustness, reliability, and controllability** of AI systems through rigorous testing.
Responsibility King (1996), Myers & Venable (2014)	**Be Responsible** EU HLEG (2019), BAAI (2019), Floridi et al. (2018), Shahriari & Shahriari (2017), McNamara et al. (2018), House of Lords (2018)	Researchers and developers should **consider potential ethical, legal, and social impacts and risks** brought in by AI.
Scientific Integrity Renaud & Zimmermann (2018)	**Be Diverse and Inclusive** EU HLEG (2019), BAAI (2019), Floridi et al. (2018), Shahriari & Shahriari (2017), McNamara et al. (2018), House of Lords (2018)	Researchers and developers of AI should **reflect diversity and inclusiveness** and benefit as many people as possible.
Property King (1996), Myers & Venable (2014)	**Open and Shared Data** EU HLEG (2019), BAAI (2019)	Researchers and developers should **make sure that there is an agreement about the ownership** of an AI. In addition, they should establish open AI platforms to avoid data/platform monopolies.
Informed Consent Myers & Venable (2014)	**Informed Consent** BAAI (2019)	Researchers and developers should ensure that users' own rights and interests are not infringed. Therefore, **the informed consent of users should be obtained**.
Trust(worthiness) Rousseau et al. (1998)	**Trustworthiness** Morley et al. (2019), Floridi et al. (2018), AI HLEG (2019)	Researchers and developers should ensure that users **perceive a high level of trust in the AI** by meeting the seven key requirements suggested by the EU HLEG.

Despite there being no fundamental theoretical IS article on the ethical dimensions of AI, we found many similarities between ethical principles from IS research and those provided in the fundamental manuscripts on the ethical dimensions of AI. In sum, the L1 papers seem to follow ethics principles from IS research, such as those for nudging (Renaud & Zimmermann, 2018), privacy (Greenaway et al., 2015), design science research (Myers & Venable, 2014), and Internet communities (King, 1996), without directly referring to them. Floridi et al. (2018) found that the already established principles for the use of AI differed only slightly and simply added the principle of explicability to their framework. We go a step further and conclude that the ethical principles established for the ethical dimensions of AI hardly differ from the existing ethical guidelines in IS. To demonstrate this, we provide an overview of IS ethics principles for researchers and the principles contained in our L1 papers in Table 3. However, these principles are of high importance, as AI, on the one hand, is constantly evolving and, therefore, needs ethical observation. On the other hand, AI may soon permeate nearly every aspect of our lives, which is different from other technologies. Furthermore, the perception of AI differs; it can be perceived either as a tool or as a moral agent. As there are many synonyms for certain ethical principles, it is important to provide aggregated ethical dimensions of AI as a starting point for further research.

However, the principles are not clearly delineated in the literature. Even though we found overall criteria for differences in the identified principles, we also found distinct overlaps in the literature. For example, Floridi et al. (2018) concluded explicability – which has been used synonymously with explainability – AI would enable the principles of beneficence, non-maleficence, justice, and autonomy. We have highlighted these overlaps in Figure 4. Achieving trustworthy AI was described as the overarching goal in three of the fundamental manuscripts (Morley et al., 2019; Floridi et al., 2018; AI HLEG, 2019) or as one of the greatest challenges (Shahriari & Shahriari, 2017), we also consider it the most important dimension that can be enabled by respecting the other principles. Trustworthiness in AI can be achieved, for example, according to Floridi et al. (2018), if the five main criteria of beneficence, non-maleficence, justice, autonomy, and explicability are fulfilled. Moreover, these criteria were discussed by all fundamental manuscripts that addressed ethical principles for AI (see Table 3). The principles in the inner part of Figure 4, in contrast, were used in these papers to describe the principles in more detail.

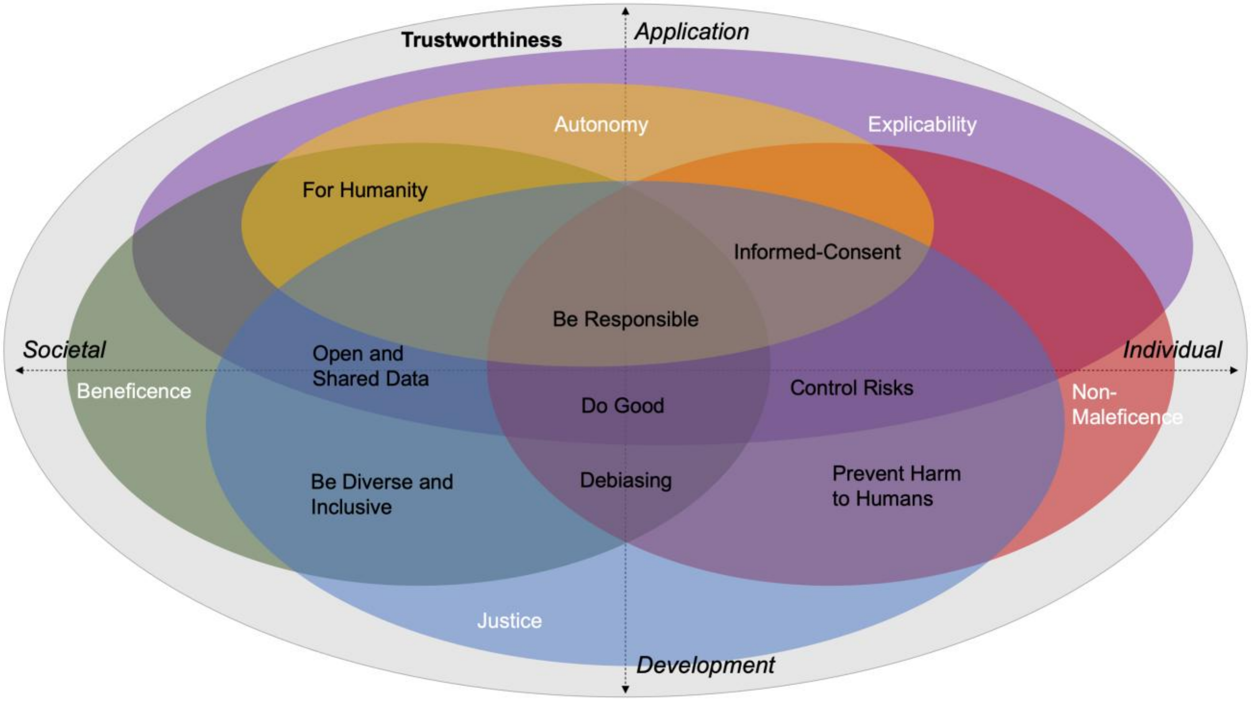


Figure 4. Classification of the Identified Ethical Principles for AI in the Dimensions Of Application, Development, Society, and Individual.

We also noticed that the ethical dimensions of AI were discussed from different perspectives. For example, some ethical principles (e.g., debiasing) refer to the development of AI-based systems and some to the application (e.g., autonomy). Also, some moral principles relate more to the impact on society (e.g., for humanity), whereas others relate more to the impact of individuals (informed consent). In Figure 4, we, therefore, classified all principles into the four dimensions “societal,” “individual,” “application,” and “development.” Even if existing ethical principles can never be unambiguously assigned to one of these dimensions, they tend to address either societal aspects or individual aspects. Even if trustworthiness can be regarded as the overriding ethical principle, subordinate principles relate either more to the applications of AI-based systems (e.g., explicability) or more to the development of AI-based systems (e.g., be diverse and inclusive).

The classified principles can also be further discussed in the context of existing literature. For example, algorithmic bias, which, according to Kordzadeh and Ghasemaghaei (2021), has not yet been investigated enough empirically, can be classified under the dimension of development and concerns both societal and individual issues. This is covered in Figure 4 with debiasing and should also be further investigated in our opinion. Phenomena such as algorithmic aversion, which was raised by Dietvorst et al. (2015, 2018), can be more closely allocated to the dimensions of individual and application, as it is related to the principle of autonomy. In contrast, algorithmic appreciation, as studied by Logg et al. (2019), can unequivocally be classified under the principle of informed consent, but also to the principle of explicability, as, for example, laypeople need to be informed of what exactly they are agreeing to when interacting with an AI-based system. The work of Leidner and Tina (2021) can rather be classified in the dimensions of individual and development, as it deals with preventing harm to humans and, thus, non-maleficence. This classification not only provides material for further discussion but also helps future research to focus on specific dimensions and explore them in more depth.

In addition to these principles of the ethical dimensions of AI for research, we identified further principles for the practical use of AI by organizations and governments. Organizations should educate and train their employees in order to improve the adaption of AI on the psychological, emotional, and technical levels (BAAI, 2019; EU HLEG, 2019; Shahriari & Shahriari, 2017). Governments should optimize employment to give full play to human advantages in order to avoid job losses and unemployment (BAAI, 2019). The Beijing AI Principles call for more cooperation, interdisciplinary work, and continuous improvement and rethinking of the principles (BAAI, 2019). Even if these aspects originally refer to governments, they can also be applied to research. Our results showed the interdisciplinary nature of research on the ethical dimensions of AI. Nevertheless, this research needs better coordination and collaboration between the different disciplines.

One question that arises is whether there are L1 papers on the ethical dimensions of AI that integrate the identified principles, guidelines, and frameworks. A clear agenda for future research on AI and ethics would also be extremely valuable. There is a lack of clear definitions and conceptualizations of what constitutes AI ethics. IS research, which otherwise addresses ethics in detail, seems disengaging and not very visible in the ecosystem of this research domain. Articles such as one by de Almeida et al. (2020) only scratch the surface of the overall discourse and offer hardly any concrete principles for the ethical dimensions of AI. Other IS articles focus more on a practical contribution rather than on a contribution to the research discourse (Martin, 2019; Robbins & Wallace, 2007). Although Porra et al. (2019) point out the importance of theoretical discourse on the ethical dimensions of AI, they do not provide concrete guidance for future research.

Therefore, we derived research questions for each ethical dimension of AI in section 5.2 to guide future IS research.

5.2 Implications for IS Research

The following implications can be derived from the interpretation of our results. First, our adapted discourse approach can be used to identify fundamental manuscripts of a current discourse based on citations and their weighting. Although we started from an IS point of view, other disciplines would find a very similar basis of L1 manuscripts in their search. Our approach provides a good starting point to identify an ecosystem of L1, L2, and L3 manuscripts.

Second, Google Scholar citations and citations in other databases are not decisive for the importance of a paper in a certain discourse, such as the ethical dimensions of AI. We found many fundamental manuscripts that had no or few citations. Other manuscripts with a high number of citations on Google Scholar or Scopus, however, could not be identified as fundamental to the considered discourse.

Third, to avoid biases, it is important that non-peer-reviewed manuscripts, conference articles, and other forms of documents are included in the search. Among the fundamental manuscripts, we found conference papers, reports, and white papers from governments and institutions. Thus, a literature search should not only focus on selected journals such as the Basket-of-Eight or a specific time period; otherwise, important papers cannot be identified.

Fourth, the discourse on the ethical dimensions of AI in IS remains fragmented and without a clear structure. So far, there are no fundamental manuscripts from IS that are directly linked to the general interdisciplinary discourse. IS literature refers to publications from the fields of philosophy and computer science as fundamental manuscripts. However, there are many similarities between the traditional ethics principles in IS research and the ethical principles of AI.

Fifth, most fundamental manuscripts on the ethical discourse in relation to AI refer to each other. However, there is no research article that links all existing principles and guidelines and discusses them in a scientifically sound manner, although Floridi et al. (2018) are very close to that. Other fundamental manuscripts, however, are not connected to other relevant papers and opened their own sub-discussions within the discourse.

Sixth, since AI technologies are constantly evolving, there cannot be universally valid and permanent principles that adhere to all ethical dimensions of AI. Existing principles and guidelines need to be continuously revised and supplemented.

5.3 Directions for Future Research

Following Pienta et al. (2020), we identified research questions and directions for IS research for each ethical dimension of AI. With these research questions, we do not claim to create an exhaustive list. Rather, we offer initial questions referring to each dimension that can be used by IS scholars as a starting point for discussion and further questions. We derived the questions from an interpretation of the future research chapters of the traditional IS literature on ethical principles and from the 12 fundamental manuscripts that we were able to identify using our manual detection method. As an example, one important question regarding the dimension of informed consent of users could be how AI can be designed by internal parties and third parties to ensure that users' rights and interests are recognized. The research questions and research directions were classified according to our identified ethics principles for developing and using AI-based systems. With related ethical themes, we provided a higher level of abstraction, which relates back to the classification in Figure 4. Figure 4 focuses primarily on the visual classification of the principles and themes in the four dimensions: societal, individual, application, and development, as well as the relationship of the principles to each other, and offers material for further scientific discourse. In accordance with Figure 4, we also show the main and tendency dimensions for the principles in a table. Table 4, in contrast to Figure 4, goes a step further and offers concrete research questions and directions for future research. The ethical themes of beneficence, non-maleficence, justice, autonomy, and explicability build on the principle of classification by Floridi et al. (2018), and the overarching principle of trustworthiness was derived from AI HLEG's (2019) discussion on trustworthy AI. The research questions and directions for IS research are shown in Table 4.

Table 4. Guiding IS Research on the Ethical Dimensions of AI by Providing Exemplary Research Questions and Directions.

Ethical Theme	Ethics principles for AI	Dimensions	Possible research directions and questions
Beneficence	Benefit Humanity	Societal, Application, & Development	**Example research questions:** • *What positive effects can be achieved for society using self-driving shuttle services in smart cities?* • *How can intelligent assistance systems be used in hospitals to relieve nurses of their workload and allow them to spend more time with their patients?* **Directions for IS research:** • Conduct design science research on new societal AI applications in healthcare or governance. • IS lecturers need to teach their students not only commercial AI applications, but also societal applications.
Non-Maleficence	Prevent Harm to Humans	Individual, Application, & Development	**Example research questions:** • *Which tasks and decision-making functionalities should not be delegated to AI-based systems to prevent harm to humans?* • *What are the design principles for AI in recruiting that help to prevent harm to applicants?* **Directions for IS research:** • Conduct quantitative research on misuse of AI applications through organizations and highlight how harm to humans can be prevented and human digital dignity can be preserved. • IS lecturers need to increase awareness of possible misuse of AI and teach how caution can be implemented into AI-based systems.

Table 4. Guiding IS Research on the Ethical Dimensions of AI by Providing Exemplary Research Questions and Directions.

Justice Explicability	Act Debiasing	Societal, Individual, & Development	**Example research questions:** • *What explanations lead to the understanding of an AI-based conversational agent by elderly people?* • *How do journalists in media organizations need to be trained to avoid data bias from an AI being used?* **Directions for IS research:** • Conduct qualitative research on mechanisms that lead to more fairness and earlier detection of biases in data used by an AI and ensure a high level of transparency, explainability (explicability), predictability, traceability, and accountability for study participants and your paper's audience. • Lecturers need to teach strategies and approaches for reducing possible discrimination (e.g., through training data) of AI-based systems.
Beneficence Non-Maleficence Justice Explicability	Do Good	Societal, Individual, Application, & Development	**Example research questions**: • *How can AI-based systems be used on social media platforms to detect and counteract fake news and misinformation?* • *What can we learn from green IS to develop green AI applications that support sustainable use cases?* **Directions for IS research:** • Identify stakeholders such as employees or customers that could be affected by (future) AI introductions (in qualitative studies) and develop targeted applications for these groups (in design science studies). • IS lectures and seminars should not be limited to the features and capabilities of AI, such as certain machine learning or deep learning algorithms, but also teach awareness of ethics and the most important application fields for societal issues.
Explicability	Ensure Autonomy	Societal, Individual, & Application	**Research questions:** • *What functionality needs to be built into self-driving vehicles to enable manual occupant intervention?* • *How can remote organizations mitigate algorithmic control to provide more autonomy for their employees?* **Guidance for IS research:** • Conduct behavioral research on the effects of algorithmic control, algorithmic aversion, and algorithmic appreciation on employees and provide guidelines to mitigate negative effects. • IS lecturers need to teach how students can design AI-based systems that provide a high level of user control.
Non-Maleficence Justice Explicability	Control Risks	Individual, Development	**Research questions:** • *What precautions can organizations take to provide the highest possible level of security and prevent cyberattacks on an AI-based system?* • *Which robustness checks do emergency management organizations need to apply before using an AI-based system in crisis communication?* **Guidance for IS research:** • Before applying an AI-based system in a study or in practice, conduct a risk analysis to control the maturity, robustness, reliability, and controllability of AI systems. • Modules for controlling AI risks and cyber threats need to be created in study programs at universities and technical colleges.
Beneficence Non-Maleficence	Be Responsible	Societal, Individual, & Application	**Research questions:** • *Which preconditions need to be established by institutions before applying AI-based systems in education?* • *How can uncertainty among employees in organizations*

Table 4. Guiding IS Research on the Ethical Dimensions of AI by Providing Exemplary Research Questions and Directions.

Justice Autonomy Explicability			*be mitigated before and after an AI-based change process?* **Guidance for IS research:** • Conduct qualitative and quantitative research on the social effects of the introduction of AI-based systems and provide guidance on how to mitigate risks. • IS decision makers (such as professors or heads of departments) should provide supplementary lectures and seminars on legal and social responsibility in organizations to establish grounding knowledge among students.
Beneficence Justice	Be Diverse and Inclusive	Societal, Development	**Research questions:** • *How can an AI-based system be used by governments to distribute information in a wide range of languages in order to better include minority groups in society?* • *How does an AI-based system need to be designed to enable people with speech disorders to comfortably communicate with non-disabled people?* **Guidance for IS research:** • Conduct qualitative and design science research on how AI needs to be trained to reflect diversity and inclusiveness and benefit as many people as possible, e.g., by recruiting study participants of minority groups or by designing targeted AI solutions. • Lecturers need to teach their students how they can reflect on diversity and inclusiveness when designing and developing AI-based systems. In addition, lectures need to address accessibility criteria for AI-based systems.
Beneficence Justice Explicability	Open and Share Data	Societal, Application, & Development	**Research questions:** • *How can blockchain technologies be used to share the ownership of AI-based systems in order to avoid data monopolies?* • *How can digital nudging be applied to engage researchers and developers of AI to establish open AI platforms and share AI-related data?* **Guidance for IS research:** • When researching or developing AI-based systems, build on open-source solutions and share your data. • Lecturers need to teach open access and open-source AI frameworks instead of teaching commercial solutions to increase awareness of open data and open science.
Non-Maleficence Justice Autonomy Explicability	Obtain Informed Consent	Individual, Application	**Research questions:** • *How can AI policies of third parties be intertwined with informed consent for AI use?* • *Which criteria do hospitals need to include in their consent forms for applying AI-based systems for supporting treatment decisions?* **Guidance for IS research:** • When conducting qualitative or quantitative research on AI, ensure that informed consent of participants is obtained and develop templates for informed consent for AI applications. • Lecturers need to teach students how to design consent forms for applying AI-based systems.
Trustworthiness	Achieve Trustworthiness	Societal, Individual, Application, & Development	**Research questions:** • *How can ethical principles be applied to conversational agents to increase the trustworthiness of public institutions during crisis events?* • *How can the seven key requirements suggested by the AI HLEG be implemented in AI-based systems to achieve a high level of trust in AI?* **Guidance for IS research:**

Table 4. Guiding IS Research on the Ethical Dimensions of AI by Providing Exemplary Research Questions and Directions.

			• Conduct interdisciplinary research on how ethical cues can be implemented in AI-based systems (such as conversational agents) to achieve a high level of trust in the system. • As trustworthiness is an overarching principle for ethical AI, lecturers need to establish courses on how to increase trust in AI-based systems.

5.4 Limitations

There are some limitations to our work. Overall, we mainly analyzed 12 manuscripts in detail. Since our primary goal was to identify fundamental manuscripts on the ethical dimensions of AI, we did not further examine L2 and L3 papers. There is also a chance that there are a few more L1 papers pertaining to the current discourse that we could not identify with our approach. For AI, there are many synonyms, and our initial keyword search was limited to rather broad search terms.

We used our adapted discourse approach for the first time and determined the threshold for the manuscripts that we classified as fundamental by visualizing the distribution curve of all calculated values. While this could lead to a small number of unknown L1 manuscripts, we were able to identify a very high percentage of L1 papers.

In addition, ethics and AI is a rapidly and constantly evolving research topic in IS research and beyond. Our work reflects the state of research from July 2020 and does not contain literature that was published after that time. In addition, we limited our initial literature search to the IS databases litbaskets.io and AISeL. Future research is needed to confirm whether we are correct in assuming that the identified L1 papers can also be considered fundamental manuscripts for other disciplines.

6 Conclusion

Every theory or emerging domain needs fundamental manuscripts marking the origin of a research field and enabling a critical discourse. For the ethical dimensions of AI, we were able to identify 12 fundamental manuscripts following an adapted discourse approach according to Larsen et al. (2019). We identified the manuscripts using a broad keyword search and a score based on weighted citations of the initially retrieved papers. We found not only journal publications, but also reports, white papers, and conference proceedings that we classified as relevant to the current discourse. None of these fundamental papers were based in IS research. Therefore, we derived concrete directions for future IS research and exemplary research questions. Nevertheless, many concepts from IS ethics research overlap with the various ethical principles of AI. Transparency, beneficence, autonomy, responsibility, justice, and scientific integrity were often attributed to ethical AI conduct. However, in IS, these principles have been examined in non-AI contexts such as nudging, research on privacy issues, or virtual collaboration for decades. Therefore, we derived the ethical dimensions of AI based on IS ethics principles and the ethical principles for AI in order to guide researchers and developers.

When carrying out research on AI, we recommend following the depicted principles of AI ethics. Our research agenda in Table 4 could serve as a starting point for this. As an interdisciplinary discipline, IS could provide a valuable L1 manuscript, synthesizing and extending the existing principles and frameworks not only for the IS community, but also for related disciplines such as economics, social science, computer science, cognitive science, and psychology. Future research should refer to and critically examine the fundamental manuscripts we have identified. For this, AI development, research, use, and its impact on different stakeholders should be considered more closely by IS scholars.

Furthermore, the IS community has the potential to contribute additional relevant key artifacts. It is especially important to increase the number of peer-reviewed research articles and to ensure that the fundamental manuscripts are not limited to government or corporate documents. IS research could utilize its fundamental knowledge on normative ethics that has already been gathered to discuss the ethical dimensions of AI in more detail. IS scholars could use previous knowledge, for example, from the fields of nudging (Renaud & Zimmermann, 2018), from research on ethics in Internet communities (King, 1996), or from research on privacy issues (Greenaway et al., 2015). Thus, future IS research could produce further fundamental papers that provide guidance for scholars of different disciplines, considering the 12 fundamental manuscripts we identified in this article. In sum, there is a lack of a general theory that explains the complex ethical dimensions of AI. Based on the identified fundamental manuscripts, IS scholars could derive such a theory.

Another important direction for future research is to further identify the ecosystem of the current discourse. Further research could uncover L2 and L3 manuscripts and their connections to L1 papers. To this end, the discourse approach of Larsen et al. (2019) could be continued.

Furthermore, there is a lack of frameworks to guide the ethical management of AI in profit and non-profit organizations. Here, again, the IS community could draw on its previous knowledge in the areas of IT strategy and the management of digital processes to create a scientific foundation. When AI is applied, it usually impacts the environment, and therefore, people and societies. If, for example, AI is used by NGOs or media organizations, the effects on society and people need to be examined more closely.

As a supplementary direction, the ethical dimensions of AI should be further investigated at a detailed level. Future research should investigate how and whether people are influenced by AI that behaves unethically. In experiments and field studies, preventive measures could be derived to prevent unethical behavior and negative effects on society and individuals.

Overall, it can be concluded that IS research on the ethical dimensions of AI is still in its infancy. Nevertheless, based on the existing knowledge on (computer) ethics in IS, there is great potential for future research, which should be exploited.

References

Abbasi, A., Jingjing, L., CLifford, G., & Taylor, H. (2018). *Make "fairness by design" part of machine learning*. Harvard Business Review. Retrieved from https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning

Ahsen, M. E., Ayvaci, M. U. S., & Raghunathan, S. (2019). When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis. *Information Systems Research*, *30*(1), 97–116.

Almeida, P., Santos, C., and Farias, J. S. (2020). Artificial intelligence regulation: A meta-framework for formulation and governance. In *Proceedings of the 53rd Hawaii International Conference on System Sciences* (pp. 5257–5266).

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183–189.

Aversa, P., Cabantous, L., & Haefliger, S. (2018). When decision support systems fail: Insights for strategic information systems from Formula 1. *Journal of Strategic Information Systems*, *27*(3), 221–236.

BAAI. (2019). *Beijing AI principles*. BAAI. Retrieved from https://www.baai.ac.cn/news/beijing-ai-principles-en.html

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Benbya, H., Pachidi, S., & Jarvenpaa, S. L. (2021). Special issue editorial: Artificial intelligence in organizations: Implications for information systems research. *Journal of the Association for Information Systems*, *22*(2), 281–303.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433–1450.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—Algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, *63*(1), 55–68.

Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Springer International Publishing.

Boell, S., & Wang, B. (2019). www.litbaskets.io, an IT artifact supporting exploratory literature searches for information systems research. In *Proceedings of the Australasian Conference on Information Systems 2019*.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (Vol. 316, pp. 316–334). Cambridge University Press.

Brendel, A. B., Mirbabaie, M., Lembcke, T. B., and Hofeditz, L. (2021). Ethical management of artificial intelligence, sustainability. *Sustainability 13*(4), 1–18.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530–1534.

Carvalho, A., Levitt, A., Levitt, S., Khaddam, E., & Benamati, J. (2019). Off-the-shelf artificial intelligence technologies for sentiment and emotion analysis: A tutorial on using IBM natural language processing. *Communications of the Association for Information Systems*, *44*(1), 918–943.

Chatterjee, S., Sarker, S., & Fuller, M., (2009). A deontological approach to designing ethical collaboration. *Journal of the Association for Information Systems*, *10*(10), 138–169.

Coppersmith, C. W. F. (2019, April 10). *Autonomous weapons need autonomous lawyers*. The Reporter. Retrieved from https://reporter.dodlive.mil/2019/04/autonomous-weapons_law/

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, *35*(8), 982–1003.

de Almeida, P. G. R., dos Santos, C. D., & Silva Farias, J. (2020). Artificial intelligence regulation: A meta-framework for formulation and governance. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

Dias, M., Pan, S., & Tim, Y. (2019). Knowledge embodiment of human and machine interactions: Robotic-process-automation at the Finland government. In *Proceedings of the 27th European Conference on Information Systems*.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology*, *144*(1), 114–126.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

Ess, C. (2009). Floridi’s philosophy of information and information ethics: Current perspectives, future directions. *The Information Society*, *25*(3), 159–168.

Etzioni, A., & Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology*, *18*(2), 149–156.

EU HLEG. (2019). *Ethics guidelines for trustworthy AI*. European Commission. Retrieved from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, *132*, 138–161.

Floridi, L., & Cowls, J. (2019). *A unified framework of five principles for AI in society*. Harvard Data Science Review. Retrieved from https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347.

Greenaway, K. E., Chan, Y. E., & Crossler, R. E. (2015). Company information privacy orientation: A conceptual framework: Company information privacy orientation. *Information Systems Journal*, *25*(6), 579–606.

Gunning, D. (2017). *Explainable artificial intelligence* (XAI) (pp. 1–17) [Distribution Statement “A”]. DAPPA. Retrieved from https://sites.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Harrington, S. J. (1996). The effect of codes of ethics and personal denial of responsibility on computer abuse judgments and intentions. *MIS Quarterly*, *20*(3), 257.

Hofeditz, L., Mirbabaie, M., Holstein, J., & Stieglitz, S. (2021). Do you trust an AI-journalist? A credibility analysis of news content with AI-authorship. In *Proceedings of the European Conference on Information Systems 2021*.

Hofeditz, L., Mirbabaie, M., Luther, A., Mauth, R., & Rentemeister, I. (2022). Ethics guidelines for using AI-based algorithms in recruiting: Learnings from a systematic literature review. In *Proceedings of the Hawaii International Conference on System Sciences*.

Horton, H. (2016, March 24). *Microsoft deletes “teen girl” AI after it became a Hitler-loving sex robot within 24 hours*. The Telegraph. Retrieved from https://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/

House of Lords. (2018). *AI in the UK: Ready, willing and able?* Authority of the House of Lords. Retrieved from https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

Kalenka, S., & Jennings, N. R. (1999). *Socially responsible decision making by autonomous agents*. Cognition, Agency and Rationality (pp. 135–149). Springer.

Kallman, E. A. (1992). Developing a code for ethical computer use. *Journal of Systems and Software*, *17*(1), 69–74.

Kawaguchi, K. (2021). When will workers follow an algorithm? A field experiment with a retail business. *Management Science*, *67*(3), 1670–1695.

King, S. A. (1996). Researching internet communities: Proposed ethical guidelines for the reporting of results. *The Information Society*, *12*(2), 119–128.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 14–29.

Kloör, B., Monhof, M., Beverungen, D., & Braäer, S. (2018). Design and evaluation of a model-driven decision support system for repurposing electric vehicle batteries. *European Journal of Information Systems*, *27*(2), 887–927.

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388-409.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engeneering*. IOS Press.

Larsen, K. R., Hovorka, D. S., Dennis, A. R., & West, J. D. (2019). Understanding the elephant: The discourse approach to boundary identification and corpus construction for theory review articles. *Journal of the Association for Information Systems*, *20*(7), 887–927.

Leidner, D. E., & Tona, O. (2021). The CARE theory of dignity amid personal data digitalization. *MIS Quarterly*, *45*(1), 343–370.

Leonard, L., & Cronan, T. (2001). Illegal, inappropriate, and unethical behavior in an information technology context: A study to explain influences. *Journal of the Association for Information Systems*, *1*(1), 1–31.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. *Robotics and well-being* (pp. 113–133).

Mann, G., & O’Neil, C. (2016). *Hiring algorithms are not neutral*. Harvard Business Review. Retrieved from https://hbr.org/2016/12/hiring-algorithms-are-not-neutral#:~:text=Don't%20let%20the%20software%20screen%20out%20good%20candidates.&text=More%20and%20more%2C%20human%20resources,pool%20of%20potential%20job%20candidates

Martin, K. (2019). Designing ethical algorithms. *MIS Quarterly Executive, 18*(2), 129–142.

Mayer, A., Haimerl, A., Strich, F., & Fiedler, M. (2021). How corporations encourage the implementation of AI ethics. In *Proceedings of the 29th European Conference on Information Systems*.

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM’s code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2018* (pp. 729–733).

Mendling, J., Decker, G., Reijers, H. A., Hull, R., & Weber, I. (2018). How do machine learning, robotic process automation, and blockchains affect the human factor in business process management? *Communications of the Association for Information Systems*, *43*(1), 297–320.

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*.

Mirbabaie, M., Hofeditz, L., Frick, N. R. J., & Stieglitz, S. (2021a). *Artificial intelligence in hospitals: Providing a status quo of ethical considerations in academia to guide future research*. AI & Society. Retrieved from https://link.springer.com/article/10.1007/s00146-021-01239-4

Mirbabaie, M., Stieglitz, S., Brünker, F., Hofeditz, L., Ross, B., & Frick, N. R. J. (2021b). Understanding collaboration with virtual assistants – The role of social identity and the extended self. *Business & Information Systems Engineering*, *63*(1), 21–37.

Moor, J. H. (1985). What is computer ethics? *Metaphilosophy*, *16*(4), 266–275.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, *26*(4), 2141–2168.

Myers, M. D., & Venable, J. R. (2014). A set of ethical principles for design science research in information systems. *Information & Management*, *51*(6), 801–809.

Porra, J., Lacity, M., & Parks, M. S. (2019). Can computer-based human-likeness endanger humanness? A philosophical and ethical perspective on digital assistants expressing feelings they can't have. *Information Systems Frontiers*, *22*, 533-547.

Renaud, K., & Zimmermann, V. (2018). Ethical guidelines for nudging in information security & privacy. *International Journal of Human-Computer Studies*, *120*, 22–35.

Renier, L. A., Schmid Mast, M., & Bekbergenova, A. (2021). To err is human, not algorithmic – Robust reactions to erring algorithms. *Computers in Human Behavior*, *124*, 106879.

Research and Markets. (2021). *Global digital assistant market* (2021 to 2026)—*Featuring Amazon, Apple and Baidu among others*. Research and Markets. Retrieved from https://www.globenewswire.com/news-release/2021/12/20/2355021/28124/en/Global-Digital-Assistant-Market-2021-to-2026-Featuring-Amazon-Apple-and-Baidu-Among-Others.html

Robbins, R. W., & Wallace, W. A. (2007). Decision support for ethical problem solving: A multi-agent approach. *Decision Support Systems*, *43*(4), 1571–1587.

Robin, E. (2019). *Artificial intelligence: Conflicts of interest between ethics and the needs of an adapted regulation*. Curiosity. Retrieved from https://medium.com/emmanuelle-robin/artificial-intelligence-conflicts-of-interest-between-ethics-and-the-needs-of-an-adapted-d2c1512e0bc9

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393–404.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson Education Limited.

Rzepka, C., & Berger, B. (2018). User interaction with AI-enabled systems: A systematic review of IS research. In *Thirty Ninth International Conference on Information Systems*.

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

Seppälä, A., & Mäntymäki, M. (2021). From ethical AI principles to governed AI. In *International Conference on Information Systems*.

Shahriari, K., & Shahriari, M. (2017). IEEE standard review — Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)* (pp. 197–201).

Sojer, M., Alexy, O., Kleinknecht, S., & Henkel, J. (2014). Understanding the drivers of unethical programming behavior: The inappropriate reuse of internet-accessible code. *Journal of Management Information Systems*, *31*(3), 287–325.

Stahl, B. C. (2008). Researching ethics and morality in information systems: Some guiding questions. In *Proceedings of the International Conference on Information Systems*.

Stahl, B. C. (2012). Morality, ethics, and reflection: A categorization of normative IS research, *Journal of the Association of Information Systems, 13*(8), 636–656.

Stahl, B. C., Eden, G., Jirotka, M., & Coeckelbergh, M. (2014). From computer ethics to responsible research and innovation in ICT. *Information & Management*, *51*(6), 810–818.

Stieglitz, S., Mirbabaie, M., Kroll, T., & Marx, J. (2019). “Silence” as a strategy during a corporate crisis – The case of Volkswagen’s “Dieselgate.” *Internet Research*, *29*(4), 921–939.

Teodorescu, M., Morse, L., Awwad, Y., & Kane, G. (2021). Failures of fairness in automation require a deeper understanding of human-ML augmentation. *MIS Quarterly*, *45*(3), 1483–1500.

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*(2), 447–464.

Turing, A. (1950). *Computing machinery and intelligence*. Mind.

Vakkuri, V., Kemell, K.-K., Kultanen, J., Siponen, M., & Abrahamsson, P. (2019). Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study.

Verma, S. (2019). Weapons of math destruction: How big data increases inequality and threatens democracy. *Vikalpa: The Journal for Decision Makers*, *44*(2), 97–98.

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems*, *37*.

Walsham, G. (1996). Ethical theory, codes of ethics and IS practice. *Information Systems Research*, *6*, 69–81.

Webster, J., & Watson, T. R. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, *26*(2), xiii-xxiii.

Wiener, M., Cram, W., & Benlian, A. (2021). Algorithmic control and gig workers: A legitimacy perspective of Uber drivers. *European Journal of Information Systems*, 1–23.

Yampolskiy, R. V. (2016). Taxonomy of pathways to dangerous artificial intelligence. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Yapo, A., & Weiss, J. (2018). Understanding the impact of policy, regulation and governance on mobile broadband diffusion. In *46th Hawaii International Conference on System Sciences* (pp. 2852–2861).

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*(1), 717–731.

About the Authors

Milad Mirbabaie is junior professor for Information Systems & Digital Society at Paderborn University and team leader for Sociotechnical Systems at the University of Duisburg-Essen, Germany. He studied Information Systems at the University of Hamburg and received his PhD from the University of Münster, Germany. He has published in reputable journals such as Journal of Information Technology, Business & Information Systems Engineering, Electronic Markets, Journal of Decision Systems, Internet Research, Information Systems Frontiers, International Journal of Information Management, and International Journal of Human Computer Interaction. His work focuses on Sociotechnical Systems, AI-based Systems, Social Media, Digital Work, and Crisis Management.

Alfred Benedikt Brendel is associate professor for business information systems, esp. intelligent systems and services, at the Technische Universität Dresden. Alfred holds a Doctor's degree in management science, specializing in Business Information Systems, from the University of Goettingen. His research focuses on exploring the human-like design of conversational agents and its effect on users' perception, affection, cognition, and behavior. His main areas of research are digital health, smart mobility, and digital work. His research is forthcoming or has been published in leading information systems journals, including *Journal of Information Technology*, *Journal of the Association for Information Systems*, *Information Systems Journal*, and *European Journal of Information Systems*.

Lennart Hofeditz is a research associate at the research group of Professor Stefan Stieglitz at the University of Duisburg-Essen, Germany. He studied Applied Cognitive and Media Science (M.Sc.). Now, he is a PhD candidate in Information Systems at the Department of Computer Science and Applied Cognitive Science at the University of Duisburg-Essen in Germany. In his research, he focusses on socio-technical systems and ethical issues related to the application of artificial intelligence and anthropomorphic machines in organizations. He also works in a research project funded by the German Research Foundation (DFG) on research data management and open science.

Paper 3: Design principles for conversational agents to support Emergency Management Agencies

Type (Ranking, Impact Factor)	Journal article (C, 19.96)
Status	Published
Rights and permissions	Open access
Authors	Stieglitz, S., Hofeditz, L., Brünker, F., Ehnis, C., Mirbabaie, M., & Ross, B.
Year	2022
Outlet	International Journal of Information Management (IJIM)
Permalink / DOI	https://doi.org/10.1016/j.ijinfomgt.2021.102469
Full citation	Stieglitz, S., Hofeditz, L., Brünker, F., Ehnis, C., Mirbabaie, M., & Ross, B. (2022). Design principles for conversational agents to support Emergency Management Agencies. *International Journal of Information Management*, 63, 102469. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2021.102469.

International Journal of Information Management 63 (2022) 102469

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Research Article

Design principles for conversational agents to support Emergency Management Agencies

Stefan Stieglitz[a,*], Lennart Hofeditz[a], Felix Brünker[a], Christian Ehnis[b], Milad Mirbabaie[c], Björn Ross[d]

[a] *Digital Communication and Transformation, Department of Computer Science and Applied Cognitive Science, Faculty of Engineering, University of Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany*
[b] *The University of Sydney Business School, The University of Sydney, Rm 4053, Level 4, Abercrombie Building H70, NSW 2006, Australia*
[c] *Paderborn University, Warburger Str. 100, (Q3.128), 33098 Paderborn, Germany*
[d] *University of Edinburgh, Informatics Forum, 10 Crichton St, Edinburgh EH8 9AB, UK*

ARTICLE INFO

Keywords:
Conversational agents
Chatbots
Design principles
Crisis communication

ABSTRACT

Widespread mis- and disinformation during the COVID-19 social media “infodemic” challenge the effective response of Emergency Management Agencies (EMAs). Conversational Agents (CAs) have the potential to amplify and distribute trustworthy information from EMAs to the general public in times of uncertainty. However, the structure and responsibilities of such EMAs are different in comparison to traditional commercial organizations. Consequently, Information Systems (IS) design approaches for CAs are not directly transferable to this different type of organization. Based on semi-structured interviews with practitioners from EMAs in Germany and Australia, twelve meta-requirements and five design principles for CAs for EMAs were developed. In contrast to the traditional view of CA design, social cues should be minimized. The study provides a basis to design robust CAs for EMAs.

1. Introduction

In crisis situations, people use social media alongside traditional news sources to search for information about the event or to share their experiences with friends or the public (Nabity-Grover, Cheung, & Thatcher, 2020; Stieglitz, Mirbabaie, Ross, & Neuberger, 2018). This is due to a high amount of uncertainty and ambiguity particularly in the early stages of a crisis (Mirbabaie, Bunker, Stieglitz, Marx, & Ehnis, 2020). Problems of information overload, rumors, conflicting information, and mis- or disinformation (Mirbabaie et al., 2020) can result from this behavior. Whereas misinformation means propositional content that is false but unintentional, disinformation is propositional content that is false on purpose (Mingers & Standing, 2018). Previous research showed that the virality of misinformation increases during crisis events (King & Wang, 2021).

Past crises such as massive bushfires in Australia or California (Beydoun, Dascalu, Dominey-Howes, & Sheehan, 2018), floods (Tim, Pan, Ractham, & Kaewkitipong, 2017), storms (Mirbabaie et al., 2020), terrorist events (Mirbabaie, Stieglitz, & Brünker, 2021), or the Covid-19 pandemic sparked broad discussions on various social media channels. The avalanche of information mixed with misinformation on social media was in the early phase of the COVID-19 pandemic referred to as an “Infodemic” (Zarocostas, 2020), which illustrates the issues which need to be dealt with in crisis social media communication. The uncontrolled diffusion of mis- and disinformation leads to an increased demand for reliable and up-to-date information by the general public (Elbanna, Bunker, Levine, & Sleigh, 2019). Emergency Management Agencies (EMAs) are struggling to cover the demand (Ehnis & Bunker, 2020), and thus, sophisticated solutions for filling the information gap are needed. This is also due to challenges in information exchange and management such as inaccessibility of information, inconsistent formats, inadequate information streams, a low priority of information diffusion, a difficult source identification, a media storage misalignment, unreliability or unwillingness of stakeholders (Altay & Labonte, 2014). As governmental actors (Aladwani & Dwivedi, 2018), EMAs need resources or approaches to interact with a large quantity of concerned citizens (Zhang, Fan, Yao, Hu, & Mostafavi, 2019). The social media communication pattern is a one-to-many (EMA-to-public) and

[*] Corresponding author.
E-mail addresses: stefan.stieglitz@uni-due.de (S. Stieglitz), Lennart.hofeditz@uni-due.de (L. Hofeditz), felix.bruenker@uni-due.de (F. Brünker), christian.ehnis@sydney.edu.au (C. Ehnis), milad.mirbabaie@uni-paderborn.de (M. Mirbabaie), b.ross@ed.ac.uk (B. Ross).

https://doi.org/10.1016/j.ijinfomgt.2021.102469
Received 3 November 2020; Received in revised form 18 December 2021; Accepted 29 December 2021
Available online 13 January 2022

many-to-one (public-to-EMA) communication for which robust practices and solutions still need to be developed.

One important solution space to improve crisis response and emergency management is the use of information and communication technologies and artificial intelligence (AI) (Fan, Zhang, Yahja, & Mostafavi, 2021). AI can be used not only for direct crisis response during natural disasters, but also to support long-term aims such as sustainability (Nishant, Kennedy, & Corbett, 2020). During hard to predict crisis situations such as during the Covid-19 pandemic, AI-based systems can assist managers and leaders in making effective and efficient decisions (Dwivedi et al., 2020). Thus, AI-based systems can be used to analyze crisis-relevant images and assign them to a region using semantic content classification or to assess damage to specific objects, such as bridges or roads. Furthermore, Natural Language Processing (NLP) and data mining techniques can be applied to detect and predict critical events and identify patterns based on social media data (Fan et al., 2021). Another approach of applying NLP is the use of Conversational agents (CAs) (Balakrishnan & Dwivedi, 2021a, 2021b). They can not only interact with users in natural language (McTear, Callejas, & Griol, 2016) but also provide an enjoyable user experience (Diederich, Brendel, & Lichtenberg, 2019). CAs are capable of assisting users in a variety of tasks such as answering frequently asked questions or providing ideas and inspiration at the workplace (Lembcke, Diederich, & Brendel, 2020).

In crisis communication they could solve problems such as providing real-time translation for outgoing and incoming messages via social media channels, provide location-specific information, answering frequently asked questions of citizens regarding ongoing disasters fast and accurately, or autonomously collecting and analyzing disaster relevant data (Hofeditz, Ehnis, Bunker, Brachten, & Stieglitz, 2019). CAs have already been tested to autonomously answer questions from members of the public (Ahmady & Uchida, 2020) or to coordinate spontaneous volunteers (Gerstmann, Betke, & Sackmann, 2019). First studies indicate that they can be applied to disseminate and collect information in crisis situations such as water-related crises (Tsai, Chen, & Kang, 2019) or the Covid-19 pandemic (Maniou & Veglis, 2020).

However, past systematic and comprehensive information systems (IS) research on the design of CAs mainly focused on the deployment in commercial organizations for customer support (Gnewuch, Morana, Adam, & Maedche, 2017), virtual collaborative work (Brachten, Brünker, Frick, Ross, & Stieglitz, 2020), or learning environments (Graesser, Li, & Forsyth, 2014). In contrast, EMAs facing crisis situations have unique requirements such as speed, effectiveness and efficiency (Fan et al., 2021) that directly or indirectly affect the safety of human lives. This makes it problematic to rely on knowledge about CA design that was solely developed in the context of commercial organizations and businesses. We think that existing knowledge cannot simply be adopted in an emergency management environment but needs to be carefully transferred and developed. The requirements of EMAs for structure, responsibilities and operations' management significantly differ from those of commercial organizations (Ehnis & Bunker, 2020; Hofeditz et al., 2019). Thus, it is crucial to derive specific design principles of CAs in crisis communication so that they can suit the needs of EMAs in crisis situations. We therefore aim to answer the following research question from a lens of interpretivism and a constructivism ontology (Goldkuhl, 2012):

RQ: How should conversational agents be designed to improve social media crisis communication of EMAs?

We adopted an interpretivist philosophy in order to gather empirical evidence from employees of several organizations from two countries (Australia and Germany) (Kwayu, Abubakre, & Lal, 2021). Our aim was to understand how CAs need to be designed to support EMAs in their crisis communication. For this, we needed to conduct data for further interpretation. We conducted 16 semi-structured interviews with crisis management experts in Australia and Germany. Through this research, this study will enrich knowledge about CAs in crisis situations and about fighting disasters by revealing the special requirements that EMAs have for CAs, and by comparing these against current CA design principles in IS research. Furthermore, by introducing specific design principles for CAs in crisis situations, this study provides a foundation which practitioners may use to develop more sophisticated CAs, and thereby help the fight against "Infodemics" by reducing information overload and reducing false information during large scale crisis events.

2. Literature review

2.1. Specificities of Emergency Management Agencies

EMAs are typically government organizations with the focus of minimizing the effects of crisis events. Their main premise is to save lives and to minimize damage. Such organizations do not have a profit-driven focus, but their operations are limited by the funding they are receiving. The members of EMAs can be paid professionals, such as in most city fire departments, predominantly volunteers, such as in the NSW State Emergency Service in Australia, or a mixture of both, such as the Country Fire Authority Victoria in Australia.

EMAs are hierarchically structured with clear command and control systems and practices in place (Bunker, Levine, & Woody, 2015; Gupta, Starr, Farahani, & Matinrad, 2016). EMAs are operating in two distinct modes, an operational mode, when they are responding to a crisis event, and a non-operational mode in between crisis events (Ehnis and Bunker 2020). Although cooperation between different EMAs is highly important during crisis events, prior research showed that EMAs often lack interoperability due to vastly different goals among organizations (Shareef et al., 2019). Altay & Pal (2014) therefore examined, how information diffusion can be increased by establishing trust and a high level of information quality. Following an agent-based modeling approach, they concluded that cluster leads should act as hubs and establish long-term relationships in order to facilitate and filter information between agencies. For this, humanitarian operations also need to lead complex interaction between deployed technology and humanitarian groups. Considering the complex character of humanitarian operations that arises, among crisis-related issues, from rapidly formed teams, intergroup leadership might reduce complexity and increase performance among the various subgroups (Dubey et al., 2020; Salem, Van Quaquebeke, Besiou, & Meyer, 2019).

It is well known that social media is an influential communication channel during crisis events (Tim et al., 2017). EMAs realized the value social media can provide and adopted various social media platforms to their communication portfolio to provide timely trustworthy information, counter rumors and misinformation, and provide recommendations for individual actions (Elbanna et al., 2019; Hofeditz et al., 2019). Previous research highlighted that demographic characters and ethnical groups differ in responses and behaviors during crisis events, resulting in different communication strategies for EMAs (Yuan, Li, Liu, Zhai, & Qi, 2021). While EMAs have adopted social media platforms for communication with the public, they often lack the ability to systematically track and analyze social media data (Ehnis and Bunker, 2020). However, activities such as identifying and analyzing potential emergency and crisis situations, developing coping strategies, and initiating and tracking countermeasures are crucial for successful crisis management (Mirbabaie et al., 2021). While some of these strategies overlap to some extent with the approaches used by commercial organizations to engage with their audiences, a distinction must be made between the motives of commercial and EMA organizations. In contrast to traditional commercial organizations whose actions are mainly based on their own economic needs, EMA's overriding goal is to protect people and the common good. To this end, various strategies are applied by EMAs such as (local) community management, volunteer management and research (Fischer-Preßler, Schwemmer, & Fischbach, 2019). These emergency management strategies are aligned to the prevention, preparedness, response, and recovery phases of a crisis (Wenger, 2017). Emergency management activities include the provision of reliable real-time information during and between crises. EMAs use social media to provide information to the public and only to a much lesser extent to protect their own reputation. Commercial organizations, on the other hand, focus in their crisis communication predominantly on the protection of their own reputation.

Compared to the need of providing rapid and reliable information in a public crisis event by EMAs, research has shown that for traditional commercial organizations the absence of communication is a strategy that could fulfill the organization's needs (Stieglitz, Mirbabaie, Kroll, & Marx, 2019).

2.2. *Crisis communication and technology use of Emergency Management Agencies*

During crises, the public's need for information is closely related to the crisis itself as well as the degree of individual involvement. At the same time, EMAs need information from the public, for example, to maintain supply chains during crisis events (Shareef, Dwivedi, Kumar, Hughes, & Raman (2020). Furthermore, EMAs need to adapt to the ongoing development of the situation and may change communication strategies over time. Therefore, EMAs can distribute information towards non-institutional actors (Abedin & Babar, 2018), such as members of the general public, and institutional actors, such as media organizations (Mirbabaie et al., 2020). In this context, it is important that policy makers of the involved (non-) governmental organizations do not only consider crisis response and preparedness but also pursue the prevention of potential crises as well as the reconstruction of the damaged economy (Shodhi, 2016). The planned operations still need to be communicated and coordinated between the participating parties. For example, Shodhi & Knuckles (2021) highlight the various flows of information, money, and materials among several stakeholders of a development-aid supply chain. The number of different stakeholders including different requirements emphasizes that proper information technology is pressingly needed for successful coordination and collaboration. EMAs are often information starters within the emerging communication networks during a crisis (Nabity-Grover et al., 2020), whereas individuals are often information amplifiers and information transmitters (Mirbabaie et al., 2020). While social media technologies are beneficial to support emergency management-relevant tasks (Oh, Eom, & Rao, 2015), EMAs still seem to struggle with adopting these technologies into their crisis-related operations (Ehnis & Bunker, 2020). Resources, particularly in the early stages of an event, are limited (Power & Kibell, 2017), and many tasks rely on manual processes (Ehnis & Bunker, 2020).

Social media CAs, in particular chatbots, have the potential to support EMAs with their social media activities (Hofeditz et al., 2019). However, EMAs are a subset of traditional command and control organizations, and therefore, bring together their proven organizational structures, processes, technologies, and IS (Ehnis & Bunker, 2020). Consequently, EMAs cannot just unreflectively implement chatbots which were designed for commercial organizations; there is a need to rethink and critically assess the design requirements which are necessary to successfully utilize social media chatbots in an emergency management environment. As CAs are part of the multidisciplinary perspectives of artificial intelligence, challenges and opportunities need to be addressed (Dwivedi et al., 2019).

2.3. *Conversational agents for crisis communication in Emergency Management Agencies*

For conversational technologies such as CAs, some inconsistencies exist in prior research regarding terminology being used and the corresponding meaning (Brachten, Kissmer, & Stieglitz, 2021). The term CA, in the current body of knowledge, is often seen as an umbrella which includes different types of human-computer interaction systems such as chatbots (Duan, Edwards, & Dwivedi, 2019), digital assistants, virtual assistants (Mirbabaie et al., 2021) or voice assistants (Laumer, Gubler, Racheva, & Maier, 2019). CAs are ISs which can communicate with human users by using and processing natural language (Laumer et al., 2019). They have been examined in areas such as healthcare (Denecke, Vaaheesan, & Arulnathan, 2020), education (Demetis & Lee, 2018) or customer service (Gnewuch et al., 2017). In research, the terms CA, chatbots and digital assistants are sometimes used synonymously (Gnewuch et al., 2017). Nowadays CAs can act more sophisticatedly and they are applied to several tasks and processes using machine learning (Mirbabaie et al., 2020). CAs can be embodied which means that they have an animated visual representation that engages face-to-face with users (Norman & Kirakowski, 2018). CAs are actively used to assist companies in communicating with customers and have been tested in many different cases such as medicine and education (Griol, Carbó, & Molina, 2012; Laumer et al., 2019). In a commercial context, CAs are an established technology and they have been found to be very helpful in automating tasks and communication.

However, crisis communication and EMAs have different requirements which need to be addressed separately. Thus, during most crisis events such as natural disasters (Hofeditz et al., 2019) or terrorist attacks (Gupta, Starr, Zanjirani Farahani, & Ghodsi, 2020), it is very important to receive assistance in resource allocation. In the context of crisis communication and emergency management, the literature indicates that CAs are used on various social media channels in the form of chatbots.

There are examples of prototype chatbots which provide crisis-relevant information to individuals in affected areas. To reduce the problems of rumor spreading and increase reliable information on social media, Ahmady & Uchida (2020) examined the utilization of chatbots providing earthquake-related information in Japan to foreigners. This application showed that chatbots could be used to reduce language barriers and provide reliable real-time information to a specific audience. Furthermore, Tsai et al. (2019) evaluated a CA that is connected to a crisis-related data base. They showed that the CA can help crisis-affected people by providing personnel access to crisis-related data in a flood context. This allowed individuals to follow corresponding response strategies. By this, people mitigate potential harmful information related to the individual decision-making process. Beside the problems of conflicting information, rumors, or information overload, spontaneous volunteers are often a crucial factor for saving lives during a crisis. Gerstmann et al. (2019) investigated the role of CAs for coordinating the behavior of spontaneous volunteers. The scholars emphasized the potential of CAs being applicable for individual assignment and scheduling of volunteers during a crisis. This automated coordination may reduce the work-load of EMAs in crisis situations. Regarding research about CAs and task-support showed that CAs are able to reduce the cognitive load of an individual (Brachten et al., 2020) that may lead to an improved crisis management. CAs such as social media chatbots are already applied and evaluated (Maniou & Veglis, 2020). The authors investigated a working CA that disseminate accurate, timely as well as customized information. They argue that the CA's ability of providing customized information to the public is helpful to fit the individual preferences of information selection.

However, the research on the application of bots in crisis communication by EMAs is still very young. Evaluated frameworks or established design approaches in this field do not exist at this time.

2.4. *Design principles for conversational agents in IS research*

In their essence, design principles are statements that contain information and practices that need to be embedded in the design and development of IS (Chandra, Seidel, & Gregor, 2015). They consist of relevant knowledge and decisions that need to be manifested in artefacts, methods, processes, or whole systems (Mirbabaie et al., 2020). As already described in the previous sections, CAs are particularly suitable to counter challenges related to the dissemination and collection of

Table 1
Design principles for CAs in the existing IS literature.

Design Principle	Description	Source
(1) Sociability	Provide the CA with the ability to adapt its conversation style in order to communicate in the user's preferred way.	Feine et al. (2019), Tavanapour et al. (2019), Meier et al. (2019), Radziwill & Benton (2017), Misiura & Verity (2019)
	Design the agent with appealing social cues in order to contribute to the perception of humanness, social presence and enjoyment in the interaction without fostering feelings of uncanniness.	Diederich et al. (2020), Tavanapour et al. (2019), Meier et al. (2020), Strohmann et al. (2019), Radziwill & Benton (2017)
(2) Proactive Communication	Provide the CA with the ability to use proactive messages in order to automatically notify users about changes.	Feine et al. (2020), Misiura & Verity (2019)
	Equip the agent with conversational capabilities for intent detection in order to increase its usefulness, given that the input of the user can be anticipated by the designer.	Diederich et al. (2020); Tavanapour et al. (2019), Radziwill & Benton (2017)
(3) Transparency	Provide the CA with functional transparency so that users can understand its functions and decisions.	Feine et al. (2020)
	Self-identify the agent as a machine, present exemplary capabilities and offer the possibility to get in touch with a human representative in order to manage user expectations and decrease potential feelings of uncanniness.	Diederich et al. (2020), Strohmann et al. (2019), Radziwill & Benton (2017)
(4) Flexibility	Provide the CA with conversational flexibility in order to react to changing contexts, tasks, and data requests.	Feine et al. (2020), Radziwill & Benton (2017)
(5) Usability	Provide the CA with user-friendly interactive capabilities in order to create an effective, efficient, and satisfying communication experience.	Feine et al. (2020), Meier et al. (2020)
	Guide the user in a conversation where required, foster context-specific handling of fallbacks, and iteratively extend the agent's conversational abilities from dialogue data in order to increase the agent's responsiveness.	Diederich et al. (2020); Tavanapour et al. (2019), Radziwill & Benton (2017)
(6) Error Handling	Provide the CA with the ability to handle errors of any kind and to save them for future improvements.	Feine et al. (2020), Strohmann et al. (2019), Misiura & Verity (2019), Radziwill & Benton (2017),

reliable information (Ahmady & Uchida, 2020; Tsai et al., 2019) or support EMAs in real-time crisis management (Gerstmann et al., 2019; Maniou & Veglis, 2020). In these application fields, CAs are subject to crisis-specific requirements. Thus, we need to develop design principles aiming at alleviating crisis-related issues.

Radziwill & Benton (2017) developed a high-level list of quality attributes which should be embedded in the design of a chatbot. (1) Performance, which involves the timely and robust interaction with a user. The CA should be particularly able to handle unexpected input. (2) Functionality, which includes the functions of the CA as well as the linguistic capabilities. (3) Humanity refers to the realism of the conversation and potential ability to pass the Turing Test. (4) Affect, which encompasses the emotional capabilities of the CA. (5) Ethics, which refers to security and privacy as well as cultural knowledge and practices towards the user audience. (6) Accessibility, which refers to the ability to be operated by a diverse set of users.

At their core, CAs in the crisis management sector need to provide a comprehensive and clear human-computer interaction. Subsequently, they need to apply to interaction principles (Misiura & Verity, 2019) as outlined by Molich & Nielsen (1990): The interaction should consist of simple and natural dialogue, use language which is familiar to the intended user, use simple instructions, minimize the user's memory load, be consistent, provide feedback, provide shortcuts, and have a design that prevents errors. Further research in the context of citizen participation derived distinct design principles describing that CA should provide social cues and conversational capabilities to ensure goal-oriented facilitation as well as display messages in simple and understandable language (Tavanapour, Poser, & Bittner, 2019). Likewise, Meier, Beinke, Fitte, Behne, & Teuteberg (2020) suggest that CA should meet the user's expectation to enable goal-oriented conversation. To this end, distinct input and output devices should be supported by the CA that is based on an information-focused interface. Regardless of the place of application, Strohmann, Höper, & Robra-Bissantz (2019) postulate that a VA should provide a robustness to errors and should not pretend to be human.

However, as CAs interact with their audience through natural language, which is a quasi-social interaction where information and meaning are transferred between a human actor and a technological actor, the interaction should be able to support social triggers. Feine, Gnewuch, Morana, & Maedche (2019) identified a taxonomy of verbal, visual, auditory, and invisible social cues from the literature. Cues as a form of social signals (Feine et al., 2019) show that the meaning of the communication in CA-to-user interaction is not just transferred through the text which is provided but on multiple levels of social communication. Applying the concept of social cues towards CAs in enterprise communication, Table 1 outlines design principles for CA in IS literature.

3. Material and methods

Research that matches the unquestionable need of EMAs for more automated communication and the IS literature stream of CAs is very limited. Therefore, we followed an exploratory approach to identify design principles for CAs that can be applied by EMAs to improve their crisis communication. As this qualitative research takes the perspective of an "interpretivist" ontology, we argue that individuals "*do not passively react to an external reality but, rather, impose their internal perceptions and ideals on the external world and, in so doing, actively create their realities*" (Suddaby, 2006, p.636). Thus, to obtain and understand the individual perspectives and relationships (Morgan & Smircich, 1980), we conducted 16 semi-structured interviews (Myers & Newman, 2007) with representatives of EMAs from Australia and Germany. Two trained researchers coded the transcripts of the interviews. Based on a random interview sample including 62 code segments, a reliability score for coding data of $\kappa = 0.95$ could be reached (Cohen, 1960). Based on the strength of agreement classification by Landis & Koch (1977), this score can be understood as *almost perfect* agreement.

Furthermore, information the interviews provide may be biased, and thus, the principle of triangulation is essential in terms of validity of the study. Triangulation is "used to refer to the observation of the research issue from (at least) two different points (Flick, 2004, p.193). In order to address multiple perspectives in our research issue's observations, we adopted a *multiple triangulation* approach (Denzin, 2009).

We chose Australia and Germany as two countries because of their federal structure and contrasting risk profile of different crisis events building the prerequisite for the *triangulation of data* in qualitative research. Furthermore, we conducted interviews from two different countries, at different times, in different places and from different people to further ensure proper triangulation of data that allows the transferability of our findings by not focusing on a single source (Patton, 1999). To balance out subjective influences of individuals, we also aimed for *investigator triangulation* using two different interviewers (Flick, 2004). The perspective of a researcher can have a significant influence on the entire research design (Clarke & Davison, 2020). We therefore discussed findings and coding among the individual authors' perspectives to further balance subjective influences. Regarding the *triangulation of theories,* we aggregated design principles based on various IS research perspectives as referred in Table 1. This juxtaposition ensures considering multiple perspectives on the design of CAs.

Furthermore, the researchers could get access to experts from several emergency management organizations through existing collaborations in these countries. We consulted experts that work in the area of crisis communications, social media crisis communication, intelligence, and operational response on a state level as their agencies are in charge during large scale crisis situations. The organizations we considered included EMAs that are in charge or at least involved during major crisis situations such as natural disasters (pandemics, forest fires, floods, etc.) or man-made disasters (terror attacks, oil spills, financial crises, ect.). A complete list of all interviewees can be found in the Appendix in Table 4.

For conducting the interviews, we used two interview guides (one in German and one in English) divided into six main sections. For the interview guides we considered different categorizations of crisis situations (Imran, Mitra, & Castillo, 2016; Wenger, 2017) and provided a definition of CAs (Gnewuch et al., 2017). After the introduction part, the use of social media by EMAs was queried. We asked concrete questions related to social media goals, guidelines, strategies, and types of messages that they publish during disasters. To determine the interviewees' role in crisis communication, the third section of the interview dealt with questions about concrete disaster cases. This included aspects like subjects' involvement and participation (Kamboj, Sarmah, Gupta, & Dwivedi, 2018). We focused on their practical work as EMAs, but also on their crisis communication during these events.

As a transition to the next part of the interview, the participants were asked if they knew of any chatbot activities during disasters. If not, they were asked what they generally imagined when they thought of bots and if they had ever recognized any automated accounts on social media platforms. We then asked if the subjects used CAs in their organization and if so, how they used them. To examine suitable application fields of chatbots in respective disaster phases, the fifth part of the interview emphasized the occurring problems and needs of organizations who use online communication for disaster management. Afterwards, we asked about challenges of social media emergency management. Interviewees were asked to highlight areas in which CAs could be applied, based on their knowledge of missing aspects and problems with the crisis communication. In the last interview section interview partners were asked to name the most important tasks in online communication during a disaster. Based on this, they were then asked which specific tasks CAs could take over to support the EMAs. Finally, we asked the interviewees whether they saw problems in the use of chatbots or if there were areas that should not be adopted. Overall, the approximately one-hour interviews contained 18 main questions with several subquestions.

The interviewees were recruited by email and through existing contacts via phone. They received an information sheet in advance and they were informed about the general conditions of the interview on the interview consent form, which ensured that they agreed that the interviews were recorded and notes taken. All interviews were conducted by two researchers each. We interviewed all experts at their usual workplaces and conducted the interviews when there was no acute crisis situation, so that the emotional, cognitive and motivational condition of the subjects could be described as stable. The interviews were transcribed manually.

We started analyzing our data with open coding (Glaser & Strauss, 2017). We then carried out a qualitative content analysis according to Mayring (2015) to code the data and to derive a category system. The goal of the content analysis was to identify specific requirements for chatbots that can improve the crisis communication of EMAs. Therefore, the analysis form of reduction was selected, to summarize the interview materials to the essential components and to provide appropriate categories suitable for the research questions. We created a codebook with eight coding categories including:

1. Contextual requirements of chatbots in crisis communication
2. Technical requirements of chatbots in crisis communication
3. Organizational requirements of chatbots in crisis communication
4. Legal requirements of chatbots in crisis communication
5. Reasons for EMAs to apply a chatbot for their crisis communication
6. Existing implementation approaches for chatbots in EMAs
7. Possible challenges and problems of using chatbots in an EMA
8. Reasons not to use chatbots in an EMA

After categorizing the interview data according to our codebook, we extracted meta requirements for CAs in crisis communication and management. For this, we followed Gnewuch et al. (2017).

4. Results

We found that all interviewees were very receptive to CAs and other forms of automated crisis communication and some were already using or testing the application of chatbots for their crisis communication. Our interviewees mentioned common requirements for CAs as a support in their organizations such as the ability to answer frequently asked questions in the context of disasters such as bushfires or floods (RMM). However, we found that in crisis communication there are also specific requirements for CAs to support both organizations and the public. For example, three interviewees (CL, EMI, REC) stated that CAs supporting crisis communication should actively ask users for further information about the crisis in their environment: "Then you might have a bot that might go, "Hey, your photo looks really interesting to us. We'd like to use it to help respond better. Could you please tell us when you took the photo, where you took it?" (EMI). This led us to MR1, the CA should actively ask for further information on the crisis (e.g., a fire, flood or storm) in the user's environment.

Another important requirement we identified was the reduction of social cues to a minimum. It was important to the interviewees that communication with a CA was purely functional and focused on content: "But making sure that what you're putting out is, like I said, [.] it's not confusing, and it's concise and clear" (PCB). This was mentioned especially in the context of short-term crisis events such as bushfires in Australia (PCB). This led us to MR2, social cues should be reduced to a minimum (see Table 1).

Another specific requirement for a CA in crisis communication that we identified is to label the source and how up-to-date the information is. As interviewee CL said: "This [information] is from [fire department], the official site. This [information] is the update". The information source could be linked to allow users to be directed to the source (CL, REC). This requirement was mentioned in the context of many different disaster types and led us to MR3, the CA should indicate the source and timestamp of each piece of information it provides.

It should also be clearly indicated whose opinion the CA represents: "[if] it's not labeled as a social media thing but it's an official advice from [fire department] or police or whatever then people will trust it". For this purpose, the CA must also be clearly marked as non-human. With one exception, the interviewees agreed on this point: "make sure that people do know that they're talking to a bot" (CL). This requirement was mainly mentioned in the context of fires and led us to MR4.

Another requirement (MR5) that we identified was that the CA should also clearly communicate how the user data is processed: "It's about privacy" (EMI). Since user inputs are partly used by organizations to improve their response to a crisis, the user must be informed about how they are used. These three requirements MR3, MR4 and MR5 thus

Table 2
Meta requirements derived from interviews.

Meta requirement	Interviewees
MR1: The CA should actively ask for further information on the crisis in the user's environment.	CL, EMI, REC, VSM, MAN, FFC, DRN
MR2: Social cues should be reduced to a minimum.	REC
MR3: The CA should indicate the source and timestamp of each piece of information it provides.	ASE
MR4: It should be clearly visible to users whose opinion the CA represents and that they are communicating with a CA.	CL, REC, EMS, VSM, JJN
MR5: The CA needs to clearly communicate how the user's input/data is processed.	EMI
MR6: The user should be able to input not only text, but also pictures, videos and location data.	EMS, EMI, VSM, FFS
MR7: The CA should be able to provide location-based information.	PCB, CMM, REC, FFS
MR8: The CA should be able to process multiple languages such as local languages and languages of minorities.	RMM, FFS
MR9: It should be ensured that the CA is connected to the systems and databases of the EMAs in order to retrieve information and store user inputs.	CL, VSM, FFS
MR10: It should be ensured that the CA can be accessed not only at one, but at multiple contact points.	PCB, CL, EMS, VSM, JJN, FFS
MR11: It should always be possible that the user is forwarded to a human.	CL, VSM, MAN, DRN, FFS
MR12: The CA should also be able to answer questions not directly related to the crisis.	REC, JJN, FFC, DRN

Table 3
Derivation of the design principles based on identified meta-requirements.

Design principle	Corresponding meta requirements	Description
DP1: Targeted communication in Crisis Situations	*MR1, MR2*	Provide the CA with a minimum of social cues and actively ask people for further information regarding the crisis event in order to focus on providing and distributing specific knowledge.
DP2: Special transparency during the Crisis Situation	*MR3, MR4, MR5*	For every piece of information, provide a suitable source (provided with a URL to further information) and a time stamp, explain how the user's input is processed. Furthermore, label the CA as a bot of a specific organization in order to achieve a high level of trust.
DP3: Appropriate implementation of the CAs in EMAs	*MR6, MR7, MR8*	Provide the CA with location-based information and the functionality to allow media content (text in multiple relevant languages, pictures, videos), in a possible combination with location data in order to collect more information about the crisis.
DP4: Interoperable integration of CAs among different digital platforms	*MR9, MR10*	Connect the CA to the intelligence systems of the EMAs and provide the CA platforms (such as social media platforms and an official website) in order to make sure to deliver reliable and current data and to reach as many people as possible.
DP5: Take the user seriously, also if it is not crisis related	*MR11, MR12*	Provide the CA with the functionality to forward specific requests of a user which may not be crisis related to a human encounter in order to leave no question unanswered and minimize uncertainty.

aim to create trust among users through transparency.

It was also very important to the interviewees that users could not only enter text as input, but also pictures, videos and spatial information. One interviewee, as an example, stated that it would be very helpful if "[.] there is somebody there with a phone or whatever posting a video [.]" (REC) he or she could send it through a CA. Location-based information was also considered necessary as an input during fires and floods, because EMAs need it to be able to send targeted messages regarding a crisis. This led us to MR6.

The EMAs we considered have the option of disseminating information in a local area in crisis situations via technology such as an app or SMS: "Yes, an emergency alert originally came out and it would go to hardlines and mobile phones with– Where people have an address in an area. Then it progressed to where people are in the area" (PCB). According to our interviewees, a CA should also have the ability to send this location-based information to users, because some warnings and recommendations for action only apply in certain areas while in other areas they could lead to uncertainty: "A chatbot automates that. What is my fire district? It knows that based on your location. [.] this is what you need to know based on your district, because of your district and because of your fire danger rating today this is what you can do, this is what you cannot do, all of those" (EMS). Based on these requirements in the context of fires, we derived MR7 described as the CA should be able to provide location-based information.

Not only the ability to provide location-based information was mentioned frequently, but also the capability to process and respond to multiple languages. Our interviewees stated that during a bushfire in Australia a wide range of different groups of people can be affected such as tourists, immigrants or even indigenous communities which all speak different languages. According to our interviewees (RMM, EMS), an essential requirement for a CA is to understand the most common languages in the local area, because a manual answer would be too slow and too time consuming for the EMAs. That led us to MR8: the CA should be able to process multiple languages such as local languages and languages of minorities.

It was especially important to our interviewees and their organizations that the systems and databases already in use have to be connected to the CA. One interviewee said that it would be important "to provide chatbots that we would then plug in to our own systems. [.] if things are connected into each other that adds a greater value to it" (RMM). This led us to MR9.

In crisis situations, it is often difficult to reach people, because not everyone uses the same information channels. Therefore, EMAs rely on different contact points, such as different social media channels, websites or apps, to reach the largest possible percentage of the affected population. Therefore, a requirement for CAs to improve crisis communication was to place the same CA on different channels simultaneously: "It could be something that is trusted by the user who has a file or a presence on the internet, but it is actually visible in a number of different ways on different platforms but it's the same bot" (EMI). This led us to MR10 that emphasizes it should ensure that the CA can be accessed not only at one, but at multiple contact points. Both MR9 and MR10 point out a need for interoperability and integration into different systems.

Even though CAs can relieve EMAs of their work in crisis situations, there was a consensus for the context of different disaster types that there should always be the option of a user being referred to a real person (MR11).

The CA should also be able to answer questions that are not directly related to the current crisis situation (CL, REC) in order to prevent users who may need help from running into a dead end (MR12). However, in such cases, according to our interviewees, the contact to a human should always be offered directly: “They had to put in some pretty clear triggers for when something like that would activate a real person for them to then get onboard and to assist them and give them help” (EMS).

The summary of our meta requirements can be found in Table 2.

5. Discussion

This paper is at the interchange of emergency management and IS where practical strategies will contribute to mitigate the impact of a crisis. The study aims to answer the question of how CAs can be designed to improve crisis communication of EMAs and thus to fight pandemics. To this end, five major design principles revealing specific characteristics of CAs in the context of crisis communication during disasters were identified.

5.1. Design principles for CA in crisis management

Table 3 shows the derived design principles aligned with the identified meta requirements. For the derivation of the design principles, we followed the approach outlined by Lechler, Stoeckli, Rietsche, & Uebernickel (2019).

The first design principle, Targeted Communication (DP1), highlights the importance of providing the CA with a minimum of social cues. This may allow affected people to focus on reliable information. This DP contradicts the findings of Feine et al. (2019) who emphasize the importance of CA’s social cues for several CAs In the context of disasters, excluding social cues of a CA might lead to a lower application of stereotypes, e.g., gender stereotypes (Nass, Moon, & Green, 1997). Following, people focus on the information itself and are less biased by entrenched stereotypes. This may allow those affected by the crisis to save cognitive resources and directly convert helpful information into action. This may help EMAs to receive valuable information in order to obtain their supply chains during crisis events (Shareef et al., 2020). CAs can thus also provide important information as a basis for decision making, which according to Dwivedi et al. (2020) is one of the great potentials of AI-based systems. However, this could differ between types and phases of crises as these differ in terms of crisis communication strategies (Gupta et al., 2016). Furthermore, the CA needs to consider the EMA’s function during the crisis as those might be responsible for specified activities such as forecasting, the distribution of supplies, or the coordination with other (non) government organizations (Gupta et al., 2016).

DP2 aligns with previous IS research (Kim, Park, & Suh, 2020). Particularly in the context of transparency and AI, it is important to explain how the users’ input is processed and which source is subject to the CA’s message. This becomes evident, especially during crisis situations which are characterized by ambiguity and uncertainty (Mirbabaie et al., 2020), therefore, the CA as a transparent and trustworthy information provider is crucial for resolving these issues. Balakrishnan & Dwivedi (2021a) argue that is important to design the CA transparent in order to help the users perceive the CA intelligent and competence. Transparency by indicating sources and timeliness of a CA’s information can also help stakeholders to distinguish real news from fake news, which is often spread during crisis events (King & Wang, 2021). A next possible step could be an integrated Fake News Detector, which enables people to ask the CA whether information is factual or fake news. This could be realized via a database linked to a fact checking tool. Not only affected citizens could benefit from the implementation of a CA that follows DP2. A CA with this functionality could also be very useful for the communication and exchange of information among EMAs, as the arising trust can lead to a better diffusion of information (Altay & Pal, 2014). It is therefore highly recommended to consider CAs when developing new and appropriate strategies to deal with crises.

Furthermore, DP3 highlights the importance of location-based information in crisis situations. Providing the CA with the functionality of processing multiple input types and languages allows EMA to collect comprehensive information about the crisis. While users in commercial applications of CAs are usually not able to send information such as videos or location data, these rich information sources become essential in crisis situations (Konicek, Netek, Burian, Novakova, & Kaplan, 2020). Although our interviews mentioned this in the context of floods in Germany and bushfires in Australia, previous studies also highlighted the usefulness of location data in other countries ((Holderness & Turpin, 2015)Holderness and Turpin 2015). DP3 is not only relevant for crisis communication during natural disasters, but also for man-made disasters such as terrorist attacks, where information symmetry, completeness of information, private information about terrorist secrecy and deception are important (Gupta et al., 2020). Here, CAs could use different media types to gather and match information for EMAs. It should also be emphasized that the combination of location data and other data such as images or videos is also of great value for emergency management, since image data of destroyed roads, bridges or other buildings, for example, can be assigned to specific regions (e.g., by means of an AI-based system) (Fan et al., 2021). The complex and dynamic nature of disaster situations raises the need for supply chain agility (Dubey et al., 2020) and enhanced cooperation between subgroups (Salem et al., 2019) that can be managed by intergroup leadership. In this way, disaster relief material movements can be coordinated and organized. Taking knowledge from operations research, EMAs may use AI-based CAs for (inventory) management of relief materials or the alignment of relief workers (Balakrishnan & Dwivedi, 2021a). However, collaborative relationships between the various EMAs and relief workers are crucial as no single organization may manage the crisis by its own. This becomes apparent regarding the coordination between different types of organizations such as governmental and non-governmental organization among the supply chain (Shaheen & Azadegan, 2020).

In this context, DP4 highlights that the CA should be connected to the intelligence system of the EMA. This allows the organization to better process and analyze the heterogeneous data, and therefore, quickly provide reliable information. As demographic characters and ethnical groups differ in terms of their responses and general behavior during crisis events (Yuan et al., 2021), people need to access the CA through multiple contact points such as social media platforms or official websites to reach various target groups as well as the majority of the public. For example, geographical IS and social media are already used to organize local response efforts. However, this is often based on a non-organized open-source approach (Shodhi & Tang, 2014). Deploying a CA that is connected to EMAs’ systems can address challenges raised by Altay & Labonte (2014) such as inaccessibility of information, inconsistent formats, inadequate information streams, a low priority of information diffusion, a difficult source identification or a media storage misalignment by providing a natural communication channel for citizens. Furthermore, it is crucial that gathered information and resources are stored, verified, and distributed to coordinated collaboration partners. To realize this, the collaboration between different departments and EMAs needs to be improved initially, since in some cases they do not function well due to different objectives (Shareef et al., 2019). However, receiving location-based information raises further challenges on a governmental level (Aladwani & Dwivedi, 2018) as well as for EMA (Zhang et al., 2019). This highly sensitive information has to be stored, processed and provided to align to the legal requirements of the state. At the same time, the data needs to be protected against abuse.

Furthermore, DP5 emphasizes the robustness to unexpected uses of the CA. In contrast to (Cassell & Thorisson, 1999), the CA should not try to hide a lack of knowledge and force to provide no or an unsatisfying answer. The findings show that in crisis situations the CA’s replies need to be accurate, reliable and transparent. This leads to the CA having to refer to a human if he cannot give a reliable answer to the user. Relying on the system gains in importance regarding the findings of Balakrishnan & Dwivedi (2021b) conceptualizing the role of trust as a system-based belief in the context of CA interaction.

In summary, we found major similarities between the requirements of these EMAs from different countries yielding into the five design principles. This might be due to a regular exchange with EMAs from other countries (e.g., from the U.S.).

5.2. Theoretical contribution

The new design principles should be followed when developing CAs for the use of emergency management agencies during crisis situations. Previous research had already identified general design guidelines for CAs in organizations. Our paper contributes to the ongoing discussion around the use of technology in crisis situations and to the preparation of EMAs for future crisis situations is that these design principles put the specific requirements of EMAs in concrete terms. It is necessary to rethink some of the previously known principles and add important aspects.

There are certainly similarities. The ability to answer queries unrelated to the crisis instead of blindly following a script, and the ability to speak to a human when the bot fails (DP5), are not entirely new. They follow from flexibility and transparency principles identified in previous research. However, when transparency was described as an important goal of CAs (Diederich, Brendel, & Kolbe, 2020; Feine, Adam, Benke, Maedche, & Benlian, 2020), the authors meant that the agent needs to be clearly labelled as artificial, and users need to be able to understand what functionalities it offers (functional transparency). Through our interviews, it additionally became clear how crucial it is that the information offered by the agent is transparent, for example that its source is mentioned and that it is accurately dated (informational transparency, DP2).

A similarly superficial parallel that, under closer scrutiny, reveals important distinctions can be found in the descriptions of the desired communication style. Previous research identified the requirements that the CA is proactive in its communication, for example that it notifies users about information that is relevant for them instead of only taking input from them (Feine et al., 2020). The requirements identified by our interviewees go one step further. The CA should actively prompt the user to provide additional information that it might need (DP1). In combination with DP3 it further becomes clear that in the crisis context, this extends to pictures, videos and location information as well as support for multiple languages.

A key difference lies also in the alleged requirement of sociability that previous research identified. Social cues were deemed an important aspect that contributes to the perception of a CA as human-like, to social presence and to the overall enjoyment of the interaction. In contrast, our interviewees were much less enthusiastic about cues that they perceived as superfluous. The CA, it was felt, should focus on asking and providing essential information, and keep the chit-chat to a minimum (DP1).

It is already known that flexibility is an important characteristic of CAs, but previous research used this term to mean flexibility within the conversation: a good CA should not merely follow a script but it should be able to react to various situations such as unexpected requests from the user. Our interviews made clear that in the context of crisis communication, a degree of flexibility about the communication channel in which the conversation takes place and from where the CA draws its information is also crucial (DP4). This is clearly much more effort for the developers, because it requires the integration of different systems that might work with different data formats and software architectures and might not have well-defined communication interfaces.

5.3. Practical contribution

Chatbots are widely used in various areas, for example in sales and customer service, to provide a customized experience, handle complaints and answer commonly asked questions. Mobile phone users are familiar with CAs that answer questions and perform tasks such as setting reminders. Given the burden that crisis situations place on the emergency services, it does not come as a surprise that police services, fire departments and others are looking to use similar technologies in the near future.

However, our research has made it clear that there is still a fundamental gap between what current technology can offer and the vision that decision-makers in emergency service agencies have in mind for successful CAs in this area. Together, our design requirements show a vision of the CA of the future that is far more ambitious than anything that is currently on offer, and this vision has little in common with the virtual agents and chatbots of today. In this context the CA may improve the management of relief materials as well as the coordination and collaboration among the disaster relief workers that could lead to a reduced complexity of disaster situations.

The emergency CA of the future does not only respond to user-initiated conversations in the way Siri, Google Assistant and Cortana focus on answering questions and carrying out tasks after the user has initiated the conversation. Instead, it purposefully initiates conversations on its own. For example, it may approach social media users who have posted relevant content and ask them for more background information before passing this information to its owners, or it may approach social media users in a specific geographic area with relevant information or requests for information. Thus, managers in EMAs are well advised not to simply deploy traditional chatbot applications, but to adopt more intelligent systems for their crisis communications. That information could be distributed to field teams and allow a dynamic adaption of the specific leadership styles to the current hazardous situation.

It does not attempt to form an emotional connection, at least not when an acute emergency is ongoing and an efficient exchange of information is of the utmost importance. Such attempts may be better suited to longer running crises, such as the ongoing COVID-19 pandemic. Managers of EMAs need to take this into account.

In addition, the emergency CA of the future is always fully aware of the current situation by frequent updates of information sources such as databases or systems. One challenge the CA needs to face is the assimilation of emerging technologies and online communication channels. The user may be following several media channels (TV, radio, news apps) alongside social media. The CA needs to understand this context. Depending on the nature of the situation, a chatbot that is giving advice which is outdated, even if only by half an hour, may be more harmful than one that is not giving advice at all.

EMAs can therefore learn much from the interviews we examined about the opportunities that CAs offer to improve their crisis communications, but also about their challenges. CAs such as chatbots cannot simply be implemented in the same way in the context of EMAs as in other contexts. This implies that when EMAs recruit experts in assistance systems and CAs or entrust other organizations with their implementation, the developers cannot simply transfer their existing knowledge and solutions to the crisis context. Therefore, a rigorous knowledge transfer is mandatory between managers, disaster relief workers and developers to further improve collaboration resting upon shared experiences.

However, our design principles can serve as guidance to peculiarities of the crisis context that have to be addressed before CAs can be used by EMAs. We further recommend to start step by step and not by trying to take into account all of our design principles at once. Crisis communication is a sensible field where errors can make the difference between life and death. It is advisable to start with a social media chatbot first and then gradually connect the systems of the EMAs. Also, the implementation within an EMA app might be a good starting point. Subsequently, other smart-home applications such as dissemination via smart speakers (e.g., Amazon Alexa) are also conceivable in order to reach as many people as possible in crisis situations.

Of course, when doing so, the EMAs should also compare their requirements with the requirements we identified for the Australian and German EMAs that we focused on in this study, and then determine whether the identified design principles may need to be modified.

5.4. Limitations and future research directions

Our qualitative research design imposes specific limitations on our findings. We collected enough data to have a diverse sample according to our interpretative judgement and expertise in qualitative research and to be able to answer our research question (Braun & Clarke, 2021). However, the transferability to other contexts in crisis communication is constricted through the organizations and the cultural context they are situated in; transferability to a broader context needs to be carefully evaluated and further investigation (Lee & Baskerville, 2012). Also, some requirements for CAs were mentioned noticeably more often than others. Even if the quantity of statements is less relevant in qualitative research, future research should examine more closely if there is a relationship between frequently mentioned requirements and importance of these requirements.

We conducted interviews from two different countries (Australia and Germany) to ensure a broader relevance of our findings, by not focusing on a single country. However, our findings might differ in other countries and cultures. Future research could consider our findings in the context of other countries and disaster management cultures in a cross-case analysis.

When our interviewees referred to crisis events, they were usually talking about natural disasters such as fires, floods or storms. Although we interviewed experts from a variety of countries and a wide range of organizations, our findings cannot be generalized to all crisis types since our study applied an interpretationist lens to the experiences of the experts we interviewed. The requirements and applicability of our design principles might differ between disaster types. Future research should therefore examine our design principles in the context of other crises, such as the Covid-19 pandemic.

Our research highlights differences in the design of CAs for commercial organizations and EMAs. Future research needs to apply these peculiarities in other contexts and in practice building on our findings.

6. Conclusions

Current IS literature provides various perspectives for designing CA in general (e.g., Feine et al., 2019, Diederich et al., 2020). However, the crisis related requirements for CA reveal the specific need for design principles considering the perspective of a crisis (Ahmady & Uchida, 2020, Maniou & Veglis, 2020). This study reveals aggregated insights from two countries suggesting of EMAs across the globe have similar requirements regarding crisis management. This specific need is conceptualized by the derived design principles.

In summary, this study uncovered five actionable design principles representing concrete but demanding requirements for EMAs. These go far beyond the previously known requirements for general-purpose CAs used in organizations (Feine et al., 2020), but they are necessary to ensure a satisfactory crisis response (Mirbabaie et al., 2021). Arguably, these requirements also go far beyond what current technology can offer. The derived design principles form a bridge between research and practice, with clear implications for what future research can focus on to ensure that it contributes to future crises, including pandemics, being managed more effectively.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 823866.

CRediT authorship contribution statement

Stefan Stieglitz: Supervision, Funding acquisition, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Lennart Hofeditz**: Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Felix Brünker**: Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Christian Ehnis**: Conceptualization, Formal analysis, Writing – original draft. **Björn Ross**: Writing – review & editing. **Milad Mirbabaie**: Writing – review & editing, Project administration, Resources.

Declarations of interest

None.

Appendix

See Appendix Table 4.

Table 4
List of interviewees.

Pseudonym	Organization	Position	State, Country
RMM	State Level Crisis Coordination	Manager, Public Information & Warnings	VIC, AU
EMS	State Level Crisis Coordination	Social Media and Content Specialist	VIC, AU
EMI	State Level Crisis Coordination	Online Intelligence Officer	VIC, AU
DA	State Police	Information & Communications Manager	SA, AU
PCB	State Police	Emergency Manager	SA, AU
CMM	State Police	(Former) Police Officer	SA, AU
REC	State Police	(Former) Police Commissioner	SA, AU
ASE	State Emergency Service	Media Expert	NSW, AU
RTC	State Fire Department	Volunteer Coordinator	NSW, AU
CL	City Fire Service	Communications Manager	NSW, AU
FFC	City Fire Service	Communications Manager	NRW, GER
JJN	NGO	State Association Manger (Quality management and organizational development)	NRW, GER
FFS	State Fire Department	Information and communications Manager	HE, GER
VSM	State Level Crisis Virtual Support	Board Member	NRW, GER
MAN	NGO	Press Officer	NRW, GER
DRN	NGO	State Commissioner for Disaster Control	NRW, GER

References

Abedin, B., & Babar, A. (2018). Institutional vs. non-institutional use of social media during emergency response: A case of twitter in 2014 Australian Bush Fire. *Information Systems Frontiers, 20*(4), 729–740. https://doi.org/10.1007/s10796-017-9789-4

Ahmady, S. E., Uchida, O. (2020). Telegram-based chatbot application for foreign people in Japan to share disaster-related information in real-time. In: 2020 5th International Conference on Computer and Communication Systems (ICCCS), 177–181. https://doi.org/10.1109/ICCCS49078.2020.9118510.

Aladwani, A. M., & Dwivedi, Y. K. (2018). Towards a theory of SocioCitizenry: Quality anticipation, trust configuration, and approved adaptation of governmental social media. *International Journal of Information Management, 43*, 261–272, 10.1016/j.ijinfomgt.2018.08.009.

Altay, N., & Labonte, M. (2014). Challenges in humanitarian information management and exchange: Evidence from Haiti. *Disasters, 38*(S1). https://doi.org/10.1111/disa.12052

Altay, N., & Pal, R. (2014). Information diffusion among agents: Implications for humanitarian operations. *Production and Operations Management, 23*(6), 1015–1027. https://doi.org/10.1111/poms.12102

Balakrishnan, J., & Dwivedi, Y. K. (2021aaa). Conversational commerce: Entering the next stage of AI-powered digital assistants. *Annals of Operations Research.* https://doi.org/10.1007/s10479-021-04049-5

Balakrishnan, J., & Dwivedi, Y. K. (2021bbb). Role of cognitive absorptionin building user trust and experience. *Psychology & Marketing, 38*, 643–668. https://doi.org./10.1002/mar.21462.

Beydoun, G., Dascalu, S., Dominey-Howes, D., & Sheehan, A. (2018). Disaster management and information systems: Insights to emerging challenges. *Information Systems Frontiers, 20*(4), 649–652. https://doi.org/10.1007/s10796-018-9871-6

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*, 187–207 (2020). https://doi.org/10/gg3j8k.

Brachten, F., Kissmer, T., & Stieglitz, S. (2021). The acceptance of chatbots in an enterprise context – A survey study. *International Journal of Information Management, 60*(June), Article 102375. https://doi.org/10.1016/j.ijinfomgt.2021.102375.

Braun, V., & Clarke, V. (2021). To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health, 13*(2), 201–216. https://doi.org/10.1080/2159676X.2019.1704846

Bunker, D., Levine, L., & Woody, C. (2015). Repertoires of collaboration for common operating pictures of disasters and extreme events. *Information Systems Frontiers, 17*(1), 51–65.

Cassell, J., & Thorisson, K. R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence, 13*(4–5), 519–538. https://doi.org/10.1080/088395199117360

Chandra, L., Seidel, S., & Gregor, S. (2015). Prescriptive knowledge in IS research: Conceptualizing design principles in terms of materiality, action, and boundary conditions. In: 2015 48th Hawaii International Conference on System Sciences, 4039–4048. ⟨https://doi.org/10.1109/HICSS.2015.485⟩.

Clarke, R., & Davison, R. M. (2020). Research perspectives: Through whose eyes? The critical concept of researcher perspectives. *Journal of Association for Information Systems, 21*(2), 483–501. ⟨https://10.0.69.41/1jais.00609⟩.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Demetis, D. S., & Lee, A. S. (2018). When humans using the IT artifact becomes IT using the human artifact. *Journal of the Association for Information Systems, 19*(10), 929–952. https://doi.org/10.17705/1jais.00514 (ABI/INFORM Collection).

Denecke, K., Vaaheesan, S., & Arulnathan, A. (2020). A mental health chatbot for regulating emotions (SERMO)—Concept and usability test. *IEEE Transactions on Emerging Topics in Computing, 99*(1). https://doi.org/10.1109/TETC.2020.2974478

Denzin, N. K. (2009). *The research Act: A theoretical introduction to sociological methods*. New York: Routledge. https://doi.org/10.4324/9781315134543

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2020). Designing anthropomorphic enterprise conversational agents. *Business & Information Systems Engineering, 62*(3), 193–209. https://doi.org/10/gg3j7n.

Diederich, S., Brendel, A. B., & Lichtenberg, S. (2019). Design for Fast Request Fulfillment or Natural Interaction? Insights from an Experiment with a Conversational Agent. *European Conference on Information Systems*. Stockholm, Sweden: AISel.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – Evolution, challenges and research agenda. *International Journal of Information Management, 48*, 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Dubey, R., Bryde, D. J., Foropon, C., Tiwari, M., Dwivedi, Y., & Schiffling, S. (2020). An investigation of information alignment and collaboration as complements to supply chain agility in humanitarian supply chain. *International Journal of Production Research, 59*(5), 1586–1605. https://doi.org/10.1080/00207543.2020.1865583

Dwivedi, Y. K., Hughes, D. L., Coombs, C., Constantiou, I., Duan, Y., Edwards, J. S., … Upadhyay, N. (2020). Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life. *International Journal of Information Management, 55*(July), Article 102211. https://doi.org/10.1016/j.ijinfomgt.2020.102211

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., & Williams, M. D. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, Article 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Ehnis, C., & Bunker, D. (2020). Repertoires of collaboration: Incorporation of social media help requests into the common operating picture. *Behaviour & Information Technology, 39*(3), 343–359. https://doi.org/10/ggqfx6.

Elbanna, A., Bunker, D., Levine, L., & Sleigh, A. (2019). Emergency management in the changing world of social media: Framing the research agenda with the stakeholders through engaged scholarship. *International Journal of Information Management, 47*, 112–120. https://doi.org/10.1016/j.ijinfomgt.2019.01.011

Fan, C., Zhang, C., Yahja, A., & Mostafavi, A. (2021). Disaster City Digital Twin: A vision for integrating artificial and human intelligence for disaster management. *International Journal of Information Management, 56*(March 2019), Article 102049. https://doi.org/10.1016/j.ijinfomgt.2019.102049

Feine, J., Adam, M., Benke, I., Maedche, A., Benlian, A. (2020). Exploring design principles for enterprise chatbots: An analytic hierarchy process study. In: Proceedings of the 15th International Conference on Design Science Research in Information Systems and Technology (DESRIST). Conference on Design Science Research in Information Systems and Technolog, Kristiansand.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies, 132*, 138–161. https://doi.org/10/ggdtg5.

Fischer-Preßler, D., Schwemmer, C., & Fischbach, K. (2019). Collective sense-making in times of crisis: Connecting terror management theory with Twitter user reactions to the Berlin terrorist attack. *Computers in Human Behavior, 100*, 138–151. https://doi.org/10.1016/j.chb.2019.05.012

Flick, U. (2004). Triangulation in qualitative research. In: Flick, U., von Kardorff, E., Steinke, I., (Eds.), A companion to qualitative research. pp. 178–184.

Gerstmann, S., Betke, H., Sackmann, S. (2019). Towards automated individual communication for coordination of spontaneous volunteers. In: Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2019). Valencia, Spain.

Glaser, B., & Strauss, A. (2017). *Discovery of grounded theory: Strategies for qualitative research*. New York, USA: Routledge.

Gnewuch, U., Morana, S., Adam, M., Maedche, A. (2017). Towards Designing Cooperative and Social Conversational Agents for Customer Service. In: Proceedings of the Thirty Eighth International Conference on Information Systems (ICIS 2017).

Goldkuhl, G. (2012). Pragmatism vs interpretivism in qualitative information systems research. *European Journal of Information Systems, 21*(2), 135–146. https://doi.org/10.1057/ejis.2011.54

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science, 23*(5), 374–380. https://doi.org/10.1177/0963721414540680

Griol, D., Carbó, J., & Molina, J. M. (2012). Optimizing dialog strategies for conversational agents interacting in AmI environments. In P. Novais, K. Hallenborg, D. I. Tapia, & J. M. C. Rodríguez (Eds.), *Ambient intelligence—Software and applications, 153* pp. 93–100). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28783-1_12.

Gupta, S., Starr, M. K., Farahani, R. Z., & Matinrad, N. (2016). Disaster management from a POM perspective: Mapping a new domain. *Production and Operations Management, 25*(10), 1611–1637. https://doi.org/10.1111/poms.12591

Gupta, S., Starr, M. K., Zanjirani Farahani, R., & Ghodsi, M. M. (2020). Prevention of terrorism–An assessment of prior POM work and future potentials. *Production and Operations Management, 29*(7), 1789–1815. https://doi.org/10.1111/poms.13192

Hofeditz, L., Ehnis, C., Bunker, D., Brachten, F., Stieglitz, S. (2019). Meaningful use of social bots? Possible applications in crisis communication during disasters. In: Proceedings of the 27th European Conference on Information Systems (ECIS), 17.

Imran, M., Mitra, P., Castillo, C. (2016). Twitter as a lifeline: Human-annotated Twitter Corpora for NLP of crisis-related messages. ⟨http://arxiv.org/abs/1605.05894⟩.

Holderness, T., & Turpin, E. (2015). From Social Media to GeoSocial Intelligence: Crowdsourcing Civic Co-management for Flood Response in Jakarta, Indonesia. *Social Media for Government Services*, 115–133. https://doi.org/10.1007/978-3-319-27237-5_6

Kamboj, S., Sarmah, B., Gupta, S., & Dwivedi, Y. (2018). Examining branding co-creation in brand communities on social media: Applying the paradigm of Stimulus-Organism-Response. *International Journal of Information Management, 39*, 169–185. https://doi.org/10.1016/j.ijinfomgt.2017.12.001

Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems, 134*, Article 113302. https://doi.org/10.1016/j.dss.2020.113302

King, K. K., & Wang, B. (2021). Diffusion of real versus misinformation during a crisis event: A big data-driven approach. *International Journal of Information Management*, Article 102390. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2021.102390.

Konicek, J., Netek, R., Burian, T., Novakova, T., Kaplan, J. (2020). Non-spatial data towards spatially located news about COVID-19: A semi-automated aggregator of pandemic data from (Social) media within the Olomouc Region, Czechia. Data, 5(3), 76. https://doi.org/10.3390/data5030076.

Kwayu, S., Abubakre, M., & Lal, B. (2021). The influence of informal social media practices on knowledge sharing and work processes within organizations. *International Journal of Information Management, 58*, Article 102280. https://doi.org/10.1016/j.ijinfomgt.2020.102280

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Laumer, S., Gubler, F. T., Racheva, A., Maier, C. (2019). Use cases for conversational agents: An interview-based study. In: A. C. on I. Systems (AMCIS) (Ed.), Proceedings of the 2019 Americas Conference on Information Systems.

Lechler, R., Stoeckli, E., Rietsche, R., Uebernickel, F. (2019). Looking beneath the tip of the iceberg: The two-sided nature of chatbots and their roles for digital feedback exchange. In: Proceedings of the 27th European Conference on Information Systems (ECIS), 18.

Lee, A. S., & Baskerville, R. L. (2012). Conceptualizing generalizability: New contributions and a reply. *MIS Quarterly, 36*(3), 749–761.

Lembcke, T.-B., Diederich, S., Brendel, A. B. (2020). Supporting design thinking through creative and inclusive education facilitation: The case of anthropomorphic conversational agents for Persona Building. In: Twenty-Eighth European Conference on Information Systems. Marrakech, Morocco: AIS Electronic Library.

Maniou, T. A., & Veglis, A. (2020). Employing a chatbot for news dissemination during crisis: Design, implementation and evaluation. *Future Internet, 12*(7), 109. https://doi.org/10.3390/fi12070109

Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education* (pp. 365–380). Netherlands: Springer. https://doi.org/10.1007/978-94-017-9181-6_13.

McTear, M., Callejas, Z., & Griol, D. (2016). *The conversational interface*. Springer. https://doi.org/10.1007/978-3-319-32967-3 (Springer).

Meier, P., Beinke, J. H., Fitte, C., Behne, A., Teuteberg, F. (2020). Feelfit - Design and evaluation of a conversational agent to enhance health awareness. In: Proceedings of the 40th International Conference on Information Systems, ICIS 2019.

Mingers, J., & Standing, C. (2018). What is information? Toward a theory of information as objective and veridical. *Journal of Information Technology, 33*(2), 85–104. https://doi.org/10.1057/s41265-017-0038-6

Mirbabaie, M., Bunker, D., Stieglitz, S., Marx, J., & Ehnis, C. (2020). Social media in times of crisis: Learning from Hurricane Harvey for the coronavirus disease 2019 pandemic response. *Journal of Information Technology, 35*(3), 195–213. https://doi.org/10/ghcps4.

Mirbabaie, M., Stieglitz, S., & Brünker, F. (2021a). Dynamics of convergence behaviour in social media crisis communication – A complexity perspective on peoples' behaviour. *Information Technology & People (ITP)*. ISSN: 0959-3845.

Mirbabaie, M., Stieglitz, S., Brünker, F., Hofeditz, L., Ross, B., & Frick, N. R. J. (2021b). Understanding collaboration with virtual assistants – The role of social identity and the extended self. *Business and Information Systems Engineering, 63*(1), 21–37. https://doi.org/10.1007/s12599-020-00672-x

Misiura, J., Verity, A. (2019). Chatbots in the humanitarian field—Concepts, uses and shortfalls [Creative Commons Attribution-NonCommercial 3.0 Unported]. Digital Humanitarian Network (DH Network). 〈https://www.digitalhumanitarians.com/chatbots-in-the-humanitarian-field-concepts-uses-and-shortfalls/〉 (Retrieved on April 14 2021).

Molich, R., & Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM, 33*(3), 338–348. https://doi.org/10.1145/77481.77486

Morgan, G., & Smircich, L. (1980). The case for qualitative research. *Academy of Management Review, 5*, 491–500.

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*, 25.

Nabity-Grover, T., Cheung, C. M. K., & Thatcher, J. B. (2020). 'Inside out and outside in: How the COVID-19 pandemic affects self-disclosure on social media'. In *International Journal of Information Management, 55*. Elsevier, Article 102188. https://doi.org/10.1016/j.ijinfomgt.2020.102188

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology, 27*(10), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

Nishant, R., Kennedy, M., & Corbett, J. (2020). Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. *International Journal of Information Management, 53*(March), Article 102104. https://doi.org/10.1016/j.ijinfomgt.2020.102104

Norman, K. L., & Kirakowski, J. (2018)*, Vol. 1. The Wiley handbook of human computer interaction volume*. Pondicherry, India: Wiley.

Oh, O., Eom, C., & Rao, H. R. (2015). Role of social media in social change: An analysis of collective sense making during the 2011 Egypt Revolution. *Information Systems Research, 26*(1), 210–223. https://doi.org/10/f66spk.

( ) Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research, 34*(5 Pt 2), 1189–1208.

Power, R., Kibell, J. (2017). The social media intelligence analyst for emergency management. In: Proceedings of the 50th Hawaii International Conference on System Sciences. International Conference on System Sciences, Hawaii. https://doi.org/10/ghcps5.

Radziwill, N., & Benton, M. (2017). Evaluating quality of chatbots and intelligent conversational agents. *Arxiv Preprint*, 21.

Salem, M., Van Quaquebeke, N., Besiou, M., & Meyer, L. (2019). Intergroup leadership: How leaders can enhance performance of humanitarian operations. *Production and Operations Management, 28*(11), 2877–2897. https://doi.org/10.1111/poms.13085

Shaheen, I., & Azadegan, A. (2020). Friends or colleagues? Communal exchange relationships during stages of humanitarian relief. *Producation and Operations Management, 29*(10), 2828–2850. https://doi.org/10.1111/poms.13254

Shareef, M. A., Dwivedi, Y. K., Kumar, V., Hughes, D. L., & Raman, R. (2020). Sustainable supply chain for disaster management: Structural dynamics and disruptive risks. *Annals of Operations Research*, (0123456789))https://doi.org/10.1007/s10479-020-03708-3

Shareef, M. A., Dwivedi, Y. K., Mahmud, R., Wright, A., Rahman, M. M., Kizgin, H., & Rana, N. P. (2019). Disaster management in Bangladesh: Developing an effective emergency supply chain network. *Annals of Operations Research, 283*(1–2), 1463–1487. https://doi.org/10.1007/s10479-018-3081-y

Shodhi, M. S. (2016). Natural disasters, the economy and population vulnerability as a vicious cycle with exogenous hazards. *Journal of Operations Management, 45*, 101–113. https://doi.org/10.1016/j.jom.2016.05.010

Shodhi, M.S., Knuckles, J. (2021). Development- disaster recovery. Production and operations aid supply chains for economic development and post-management. 〈https://onlinelibrary.wiley.com/doi/10.1111/poms.13489〉.

Shodhi, M. S., & Tang, C. S. (2014). Buttressing supply chains against floods in Asia for humanitarian relief and economic recovery. *Production and Operations Management, 23*(6), 938–950. https://doi.org/10.1111/poms.12111

Stieglitz, S., Mirbabaie, M., Kroll, T., & Marx, J. (2019). "Silence" as a strategy during a corporate crisis – The case of Volkswagen's "Dieselgate". *Internet Research, 29*(4), 921–939. https://doi.org/10.1108/INTR-05-2018-0197

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management, 39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Strohmann, T., Höper, L., Robra-Bissantz, S. (2019). Design guidelines for creating a convincing user experience with virtual in-vehicle assistants. In: Proceedings of the 52nd Hawaii International Conference on System Sciences, 11, 54–78. https://doi.org/10.24251/hicss.2019.580.

Suddaby, R. (2006). What grounded theory is not. *Academy of Management Journal, 49*(4), 633–642.

Tavanapour, N., Poser, M., Bittner, E. A. C. (2019). Supporting the idea generation process in citizen participation - Toward an interactive system with a conversational agent as citizen participation - toward an interactive. In: Proceedings of the European Conference on Information Systems 2019. Stockholm.

Tim, Y., Pan, S. L., Ractham, P., & Kaewkitipong, L. (2017). Digitally enabled disaster response: The emergence of social media as boundary objects in a flooding disaster. *Information Systems Journal, 27*(2), 197–232.

Tsai, M.-H., Chen, J., & Kang, S.-C. (2019). Ask Diana: A keyword-based chatbot system for water-related disaster management. *Water, 11*(2), 234. https://doi.org/10.3390/w11020234

Wenger, C. (2017). The oak or the reed: How resilience theories are translated into disaster management policies. *Ecology and Society, 22*(3), art18. https://doi.org/10.5751/ES-09491-220318

Yuan, F., Li, M., Liu, R., Zhai, W., & Qi, B. (2021). Social media for enhanced understanding of disaster resilience during Hurricane Florence. *International Journal of Information Management, 57*(April 2020), Article 102289. https://doi.org/10.1016/j.ijinfomgt.2020.102289

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet, 395*(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X

Zhang, C., Fan, C., Yao, W., Hu, X., & Mostafavi, A. (2019). Social media for intelligent public information and warning in disasters: An interdisciplinary review. *International Journal of Information Management, 49*, 190–207. https://doi.org/10.1016/j.ijinfomgt.2019.04.004

Stefan Stieglitz is professor and head of the research group for Digital Communication and Transformation at the University of Duisburg-Essen, Germany. In his research, he investigates how to make use of social media data. Moreover, he analyzes user behavior and technology adaption of collaborative IS in organizational contexts. His work has been published in reputable journals such as Journal of Management Information System (JMIS), Business & Information Systems Engineering (BISE), International Journal of Social Research Methodology, and MISQe. His articles was recognized with the 'AIS Senior Scholars Best IS Publications Award' and the Stafford Beer Medal.

Lennart Hofeditz is a research associate at the research group of professor Stefan Stieglitz at the University of Duisburg-Essen, Germany. He studied Applied Cognitive and Media Science (M.Sc.). At the moment, he is a PhD candidate in Information Systems at the Department of Computer Science and Applied Cognitive Science at the University of Duisburg-Essen in Germany. In his research, he focusses on socio-technical systems and ethical issues related to the application of artificial intelligence and anthropomorphic machines in organizations. He also works in a research project funded by the German Research Foundation (DFG) on research data management and open science.

Felix Brünker is a research associate at the Department of Computer Science and Applied Cognitive Science at the University of Duisburg-Essen, Germany. He studied Applied Cognitive and Media Science at the University of Duisburg-Essen, Germany, and specialized in professional communication in electronic media/social media. Felix is a full member of the research training group "User-Centred Social Media" (DFG). His work focusses on CAs, IT identity, and Digital Work. His work has been published in reputable journals such as Electronic Markets, Business & Information Systems Engineering, Information Technology & People, or Information Systems and e-Business Management.

Christian Ehnis is an Honorary Associate at the University of Sydney Business School. He obtained his PhD from the University of Sydney. His research interests focus on how technology influences and impacts organizations and society, particularly the use of social media during emergency and disaster events. His work has been published in reputable journals such as the Journal of Information Technology, Information Systems Frontiers, and Behaviour and Information Technology.

Björn Ross is Lecturer in Computational Social Science at the University of Edinburgh School of Informatics. In his research, he uses computational methods to study social media and related technologies. A key focus of his research is to explore aspects of social media, such as misinformation, hate speech, and the malicious use of automation (bots), as well as how social media can be used effectively for the benefit of society, such as in crisis communication. His articles have been published in journals including the European Journal of Information Systems and Big Data & Society.

Milad Mirbabaie (milad.mirbabaie@uni-paderborn.de) is junior professor for Information Systems at Paderborn University and team leader for Sociotechnical Systems at the University of Duisburg-Essen, Germany. He studied Information Systems at the University of Hamburg and received his PhD from the University of Münster, Germany. He has published in reputable journals such as Journal of Information Technology, Business & Information Systems Engineering, Electronic Markets, Journal of Decision Systems, Internet Research, Information Systems Frontiers, International Journal of Information Management, and International Journal of Human Computer Interaction. His work focuses on Sociotechnical Systems, Artificial Intelligence, Social Media, CSCW, and Crisis Management.

Paper 4: Artificial intelligence in hospitals: providing a status quo of ethical considerations in academia to guide future research

Type (Ranking, Impact Factor)	Journal article (N/A, 2.6)
Status	Published
Rights and permissions	Open access
Authors	Mirbabaie, M., Hofeditz, L., Frick, N. R. J., & Stieglitz, S.
Year	2022
Outlet	AI & Society (AI&Soc)
Permalink / DOI	https://doi.org/https://doi.org/10.1007/s00146-021-01239-4
Full citation	Mirbabaie, M., Hofeditz, L., Frick, N. R. J., & Stieglitz, S. (2022). Artificial intelligence in hospitals: providing a status quo of ethical considerations in academia to guide future research. *AI & Society*, 37(4), 1361–1382. https://doi.org/https://doi.org/10.1007/s00146-021-01239-4.

AI & SOCIETY (2022) 37:1361–1382
https://doi.org/10.1007/s00146-021-01239-4

ORIGINAL ARTICLE

Artificial intelligence in hospitals: providing a status quo of ethical considerations in academia to guide future research

Milad Mirbabaie[1] · Lennart Hofeditz[2] · Nicholas R. J. Frick[2] · Stefan Stieglitz[2]

Received: 15 February 2021 / Accepted: 25 May 2021 / Published online: 28 June 2021

Abstract
The application of artificial intelligence (AI) in hospitals yields many advantages but also confronts healthcare with ethical questions and challenges. While various disciplines have conducted specific research on the ethical considerations of AI in hospitals, the literature still requires a holistic overview. By conducting a systematic discourse approach highlighted by expert interviews with healthcare specialists, we identified the status quo of interdisciplinary research in academia on ethical considerations and dimensions of AI in hospitals. We found 15 fundamental manuscripts by constructing a citation network for the ethical discourse, and we extracted actionable principles and their relationships. We provide an agenda to guide academia, framed under the principles of biomedical ethics. We provide an understanding of the current ethical discourse of AI in clinical environments, identify where further research is pressingly needed, and discuss additional research questions that should be addressed. We also guide practitioners to acknowledge AI-related benefits in hospitals and to understand the related ethical concerns.

Keywords Artificial intelligence · Ethics · Healthcare · Hospitals · Discourse approach

1 Introduction

Ethical considerations are not limited to the philosophy discipline (e.g., Ploug and Holm 2020), but are also highly relevant in healthcare and social science-related disciplines such as information systems (IS) (e.g., Wang 2020). However, current developments in artificial intelligence (AI) give rise to profound novel ethical challenges when applied in healthcare, possibly posing a threat to patients (Jain et al. 1996; Rudin 2019; Mirbabaie et al. 2021a).

The implementation of AI recently became more distributed in hospitals worldwide (Knijnenburg and Willemsen 2016; Luger and Sellen 2016; Li et al. 2019b), creating discernible benefits assisting medical experts in hospitals (Rauschert et al. 2020; Rong et al. 2020). The term AI is usually associated with human-like behavior, but it must rather be considered as a ubiquitous concept (Siau and Wang 2018). Current applications have been developed for particular tasks (e.g., Frick et al. 2019a), such as taking advantage of medical data to generate predictions or derive recommendations (Krittanawong et al. 2017; Ku et al. 2019). For example, AI monitors patients' health conditions to support healing and regeneration (Pereira et al. 2013) and assists physicians in diagnosing diseases (Mirbabaie et al. 2021b) and planning suitable treatments (e.g., De Ramón Fernández et al. 2019; Li et al. 2019a, b; López-Martínez et al. 2019). However, some AI approaches possess certain technical restrictions which can lead to diagnostic results not being transferable to other circumstances or not being comprehensible to humans, i.e. remaining a black box (Anderson and Anderson 2007; Menai 2015; Knight 2017; Burton et al. 2019; Devi et al. 2019). Scholars and practitioners are also concerned with preventing inequitable usage and unfair information

✉ Milad Mirbabaie
milad.mirbabaie@uni-paderborn.de

Lennart Hofeditz
lennart.hofeditz@uni-due.de

Nicholas R. J. Frick
nicholas.frick@uni-due.de

Stefan Stieglitz
stefan.stieglitz@uni-due.de

[1] Faculty of Business Administration and Economics, Paderborn University, Paderborn, Germany

[2] Professional Communication in Electronic Media / Social Media, University of Duisburg-Essen, Duisburg, Germany

practices (Salerno et al. 2017; Sonja et al. 2018; Libaque-Sáenz et al. 2020). Furthermore, AI still learns from medical data that is preprocessed by humans and thus might contain bias or prejudices (Kara et al. 2006; Hirschauer et al. 2015; Ploug and Holm 2020; Alami et al. 2020).

Enthusiasts claim strong reasons for the application of AI in hospitals (Ploug and Holm 2020); nevertheless, there are ominous threats possibly leading to AI becoming destructive (Arnold and Scheutz 2018). AI is a powerful but inscrutable tool unleashed with potential dubious effects for areas in which it is applied, e.g., healthcare and/or hospitals (Crawford and Calo 2016). Research on ethical considerations of AI in hospitals is no longer a mere part of science fiction but a real-world concern (Luxton 2014a, b). Despite existing studies on ethics of AI in healthcare (e.g., Alami et al. 2020; Arnold and Scheutz 2018; Ploug and Holm 2020), we argue that current research does not consider the growing significance of the topic in a diversified enough manner, but is rather narrowly focused on traditional explorations.

The current ethical discourse on AI is rather limited and usually presented in an unsystematic manner while also being conducted in separate disciplines (Brendel et al. 2021). There should instead be an increasing debate about ethical concerns (Porra et al. 2020) taking into account the multiple characteristics, principles, and dimensions of AI. Thus, our study follows a more holistic approach by identifying fundamental literature and pioneering works from diversified research domains. We aim to summarize ethical considerations into a research agenda for academia. Precisely, we intend to encourage the discourse on ethical considerations of AI in hospitals from an interdisciplinary perspective. We argue that this is of great interest to researchers and practitioners because the application of AI in hospitals is expected to increase heavily over the next decade and the impact on healthcare could be significant (Mirbabaie et al. 2021a).

Physicians still consider AI to be simple programs, tools, or algorithms that provide support in executing a certain task but they do not recognize (or even ignore) the fact that AI is capable of continuously learning and developing over time (Mitchell et al. 2018) and that it acts independently while delivering superior results compared to humans. There is an urgent demand for interdisciplinary research to comprehend the ongoing discourse on ethical considerations and dimensions of AI in hospitals and to understand the intricacies of this ever-evolving research area. By providing a holistic picture of ethical considerations and dimensions on AI in hospitals that are currently being researched, we aim to capture the current status quo and to guide pertinent future research directions. To address this urgent issue, our research is guided by the following research questions:

RQ1: *What is the current discourse in academia and what are opinions of physicians regarding ethical considerations and dimensions of artificial intelligence in hospitals?*

RQ2: *What are future directions for interdisciplinary ethical research on artificial intelligence in hospitals?*

We followed a modified discourse approach following the suggestions of Larsen et al. (2019) and identified as well as analyzed the domain ecosystem of ethical considerations and dimensions of AI in hospitals for a corpus construction. We thus performed descriptive research examining existing literature that describes the current situation (Bell 1989; Bear and Knobe 2016). In addition, we conducted semi-structured interviews with domain experts to further elaborate on and highlight related ethical challenges of AI in the clinical environment. This prescriptive approach contains implications and consequences as well as future recommendations (Bell 1989; Bear and Knobe 2016).

This paper contributes to theory by summarizing and structuring the status quo of recent research on ethical considerations and dimensions of AI in hospitals. Researchers will find the overview helpful to understand the current ethical discourse of AI in a hospital setting. To assist future investigations, we outline ethical constructs on AI in hospitals with which recent research is concerned. Furthermore, we outline an agenda explaining where further research is pressingly needed, and which questions need to be addressed. Practitioners will comprehend the differences between currently applied systems in hospitals and recent AI developments. Furthermore, medical specialists will be able to understand the extent to which AI is beneficial for clinical settings and the ways in which the stakeholders involved, i.e. physicians and patients, can benefit from its implementation. In terms of implications for society, readers will realize that AI is already used in hospitals and that its distribution continues to grow. Individuals will further understand that multiple issues regarding the application of AI in hospitals remain unaddressed.

2 Literature background

In this section, we start by explaining the concept of AI, followed by outlining illustrative examples of applications in hospitals. We then describe current ethical principles in healthcare, and finally, we illustrate ethical considerations associated with AI in hospitals.

2.1 AI applications in hospitals

Hospitals face a variety of issues that reduce the quality of care such as delayed patient flow or erroneous surgery scheduling (Ker et al. 2018; Bygstad et al. 2020). The introduction of AI might improve these types of common issues and yield sustainable advantages. This explains why medical research and practice are increasingly concerned with possible applications of AI (e.g., Bargshady et al. 2020; Jiang et al. 2017; Rauschert et al. 2020). AI is not a specific technology that is granted to a single discipline, but rather a collection of several concepts that constantly evolve (Barredo Arrieta et al. 2020). AI can generally be defined as "the ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem-solving, decision-making, and even demonstrating creativity" (Rai et al. 2019, p. iii). Simply put, AI aims to imitate human-like behavior (Krittanawong et al. 2017); however, current implementations are still far from achieving this goal (Brachten et al. 2020).

Applications of AI are rather narrowed down to a specific task (Batin et al. 2017; Frick et al. 2019b; Mirbabaie et al. 2020) but commonly generate superior results compared to humans. When integrated into the existing technical infrastructure of hospitals, AI accelerates data collection from multiple sources (Nasirian et al. 2017), provides medical experts with more accurate and timely information (Atherton et al. 2013; Preece et al. 2017; Diederich et al. 2019), tailors to the needs of patients and their treatment processes (Dilsizian and Siegel 2014) and enhances integration with other hospital IS (Serrano et al. 2020). AI continuously learns and develops over time by processing various types of medical information from multiple years of experience using divergent data sources (Mitchell et al. 2018). Conclusions are based on a larger sample size compared to those of medical professionals (Neill 2013) and AI is more likely to provide objective decisions. AI is also more likely to evaluate patients' conditions based on medical facts, as their systems do not rely on subjective impression, situations, emotions, or time of the day (Gnewuch et al. 2017; Seeber et al. 2020).

AI already supports multiple processes within hospitals. For example, AI guides patients with exercise promotion, medication adherence (Bickmore et al. 2010; King et al. 2013), chronic disease self-care management (Kimani et al. 2016), and daily diabetes routines (Shaked 2017) as well as accelerating the gathering of medical information in preparation for therapy and forwarding them to physicians (Denecke et al. 2018). In these examples, patients use AI in the form of a conversational agent (CA), intelligent systems that interact with and augment humans' abilities (Mirbabaie 2021). Interacting with CAs not only assists patients but also clinicians in the treatment of certain diseases.

AI also assists medical experts within disease diagnostics such as ectopic pregnancies (De Ramón Fernández et al. 2019), neonatal sepsis (López-Martínez et al. 2019), or coronary artery disease (Li et al. 2019a). Medical data are thereby processed, evaluated, and classified using AI algorithms to estimate probabilities and enable clinicians to detect diseases earlier, thus allowing them to treat patients more effectively. The implementation of information technologies such as AI can impact hospitals' revenue cycle management and consequent financial sustainability (Singh et al. 2021).

Even though existing endeavors provide justification for the use of AI in clinical environments, researchers and practitioners are frequently confronted with ethical questions eventually preventing possible applications due to the fear of causing unpredictable harm to patients. The discussion on autonomous driving showed that the expectations on AI can be even higher than towards human. The same could apply for the use of AI in hospitals and therefore need further examination.

2.2 Ethical principles in healthcare

Ethics is an interdisciplinary field of study and a complex concept that governs the accumulation and interplay of moral principles (Siau and Wang 2020). Moral principles describe norms for the behavior and actions of groups or individuals in a society (Nalini 2019) that guide entities (such as humans or intelligent robots) regarding what is right and wrong. Overall, it is tough to determine where ethical behaviors begin and where unethical behavior comes into play. As one approach to determine what is right and wrong, virtues can be considered. Virtue ethics is part of normative ethics and addresses the principles in which individuals believe (Siau and Wang 2020). Virtue ethics can be seen as an overarching moral principle to help make morally problematic decisions (such as which treatments should be provided in hospitals based on a diagnosis made by an AI). In this study, we therefore focused on a virtue-ethical perspective regarding AI applications in hospitals, concentrating on treatment decisions.

Research on ethical considerations in healthcare is generally divided into three fields (Page 2012): the first field focuses on ethical developments of future healthcare experts throughout their medical training (Price et al. 1998; Bore et al. 2005). The second assesses individual ethical attitudes and how they differ among medical professions (Rezler et al. 1990, 1992). The third is concerned with the evaluation of ethical principles and their applications within treatment of patients (Hebert et al. 1992; Price et al. 1998). Ethical principles in medicine can be traced back to those of the physician Hippocrates (400 BCE), on which the concept of the Hippocratic oath is rooted (Miles 2005). The Hippocratic oath was a Greek document containing ethical standards for physicians which, for example, covers protecting the privacy of patients (Fox and James 2020). Today, the majority of medical graduates swear some kind of oath that is based on the Hippocratic oath (Hulkower 2010). Since its origin, various concepts have been developed for ethical guidelines for treating patients. The principles of biomedical ethics of Beauchamp and Childress (2019) have found great acceptance in medicine. The authors define four core principles of bioethics. (1) The principle of beneficence involves the expectation that healthcare professionals act in a way that benefits patients. (2) The principle of non-maleficence aims at avoiding any harm to involved individuals, i.e., patients or physicians. (3) The principle of autonomy respects the capabilities of individuals to make independent decisions. (4) The principle of justice specifies that all patients should be treated equally (Beauchamp and Childress 2019). Treatment ethics is intentionally defined rather broadly to allow room for individual considerations and prioritizations by physicians. Besides the principles of bioethics, ongoing research and practice are increasingly shaped by associations. There are country-specific organizations like the American Medical Association (USA) or the Academy for Ethics in Medicine (Germany), which define standards for honorable behavior of physicians when treating patients and encourage the scientific discourse on ethical questions in medicine (Riddick 2003; AEM 2020; AMA 2020). Furthermore, there are overarching institutions like the European Council of Medical Orders (CEOM 2020), which promote the practice of high-quality medicine in light of the patients' needs.

Despite the existence of ethical guidelines and principles for medical professionals, the entire healthcare system is regularly confronted with new ethical considerations. A recent example from Poland demonstrates that local governments affect healthcare and affect the majority of a population. The country's constitutional court declared abortions of children with malformations to be illegal (Amnesty International 2020). Besides restricting the freedom of choice of expectant parents, practicing physicians are restrained by this law and must abide even when an alternative decision might be more appropriate. Human rights activists and the Polish opposition heavily criticized the ruling of the constitutional court, arguing that illegal abortions will rise (Walker 2020). Another example of ethical considerations is the current discussion on distributing a potential COVID-19 vaccine. In principle, it seems reasonable that vaccinations should be given in a sequence based on profession. It is suggested, for example, that people in caring jobs should receive preferential treatment. Naturally, the question arises which professions within care should be prioritized, e.g., nursing, or elderly care?

The examples presented are intended to illustrate the idea that ethical principles are not only established by medical workers but are also heavily impacted by external forces. Likewise, AI applied in healthcare needs to adjust to a continuously changing environment with frequent interruptions (Wears and Berg 2005; Menschner et al. 2011; Rosen et al. 2018), while maintaining ethical principles to ensure the well-being of patients. Thus, in our study, we use the four core principles of biomedical ethics as suggested by Beauchamp and Childress (2019) as a conceptual categorization to classify our findings. This is then used to provide a research agenda for academia to examine the ethical challenges of using AI in hospitals.

2.3 Ethical considerations of AI in hospitals

Recent AI implementations in hospitals and in healthcare in general come with a variety of ethical considerations. For example, AI is associated with bias, discrimination, opacity, and rational concerns and intentions (e.g., Arnold and Scheutz 2018; Gruson et al. 2019; Ploug and Holm 2020) as much as it is associated with transparency, trust, responsibility, liability, and explainability (e.g., Alami et al. 2020; Wang 2020). A recent study by Ploug and Holm (2020) investigated the ethical concerns of AI for medical diagnostics and treatment planning. The authors argued that patients should be able to withdraw from being evaluated by AI because a trustful relationship between physicians and patients is essential for the success of the treatment process. Furthermore, Ploug and Holm (2020) explain that there are problems regarding bias and opacity for the patient, related implications for the entire healthcare sector, and rational concerns about impacts on society. Another study by Alami et al. (2020) provides a synthesis of key challenges posed by AI. Besides technological, organizational, and economic issues, the authors also raise several ethical obstacles. For example, AI applications can be distinguished between decision-support tools and decision-making tools. AI as decision-support tools assist medical specialists with specific tasks, e.g., within the diagnostic process (e.g., De Ramón Fernández et al. 2019; López-Martínez et al. 2019). When applied as a decision-making tool, AI will derive conclusions on its own without being supervised by physicians. However, it is yet to be defined who is held responsible for AI-based decisions leading to errors in the treatment process. Another issue illustrated by Alami et al. (2020) is the potential unexplainability of algorithmic outcomes, i.e. black box, posing a high risk to patients' well-being (Knight 2017; Rudin 2019). Of course, this makes it nearly impossible to build trust in the AI's decisions, especially when patients' lives are at stake.

Compared to ethical guidelines in healthcare, there are neither standardized regulations for the application of AI in healthcare nor in hospitals. However, most healthcare systems acknowledge the rapid development of AI for medical purposes (Duan et al. 2019) causing organizations and governments to define relevant ethical frameworks. For example, the European Union has developed the "European Ethics Guidelines for Trustworthy AI" defining its recommendations for trustworthy AI and key requirements for safety and for societal and environmental well-being (EU 2020). Furthermore, the World Health Organization has explained ethical challenges for the "global development and implementation of artificial intelligence systems in healthcare" (Bærøe et al. 2020, p. 261) and continually proposes suggestions for the ethical development and usage of AI. Besides global observations of AI within healthcare, research is equally concerned with deriving ethical principles, guidelines, and frameworks. For example, Floridi et al. (2018) developed an ethical framework for a good AI society based on the four core principles of bioethics of Beauchamp and Childress (2019). The authors added a fifth dimension explicability explaining the "need to understand and hold to account the decision-making processes of AI" (Floridi et al. 2018, p. 700).

Since the authors took an initial approach to tackle ethical issues regarding AI, we extended our conceptual categorization to include the principles of biomedical ethics of Beauchamp and Childress (2019) as well as the dimension of explicability (Floridi et al. 2018) which in most research is interchangeably used for explainability. We used these two pieces of work as the foundation of this work because they have been frequently cited and are centrally concerned with ethical dimensions of AI in various domains. Additionally, we used these frameworks because one includes a clear philosophical perspective on virtue ethics and both a bioethical perspective that is applicable to treatment ethics and the context of healthcare. Even though these articles did not focus on healthcare or hospitals themselves, the discussed ethical principles have been frequently used in other articles. Despite increasing studies being conducted on ethical considerations, current approaches are mostly congruent or very alike and focus on one specific discipline or a certain abstraction level. We thus argue that future endeavors would benefit from an alternative discourse from an interdisciplinary perspective that guides pertinent research directions.

3 Research design

Ethical discourses on the impact of new technologies are usually very unsystematic, as there is often no fundamental manuscript on which to base them. Although there have been some pioneering works, which are often quoted, many parallel discourses emerge, which make little reference to each other. In addition, ethical discourses are usually conducted separately in certain disciplines. To investigate how academia can contribute to the responsible use of AI in digital health and practical health in hospitals, we identified fundamental manuscripts following adapted version of the discourse approach proposed by Larsen et al (2019). Based on this, we identified ethical principles and their relationships and highlighted these via expert interviews with hospitals physicians and other decision-makers in hospitals.

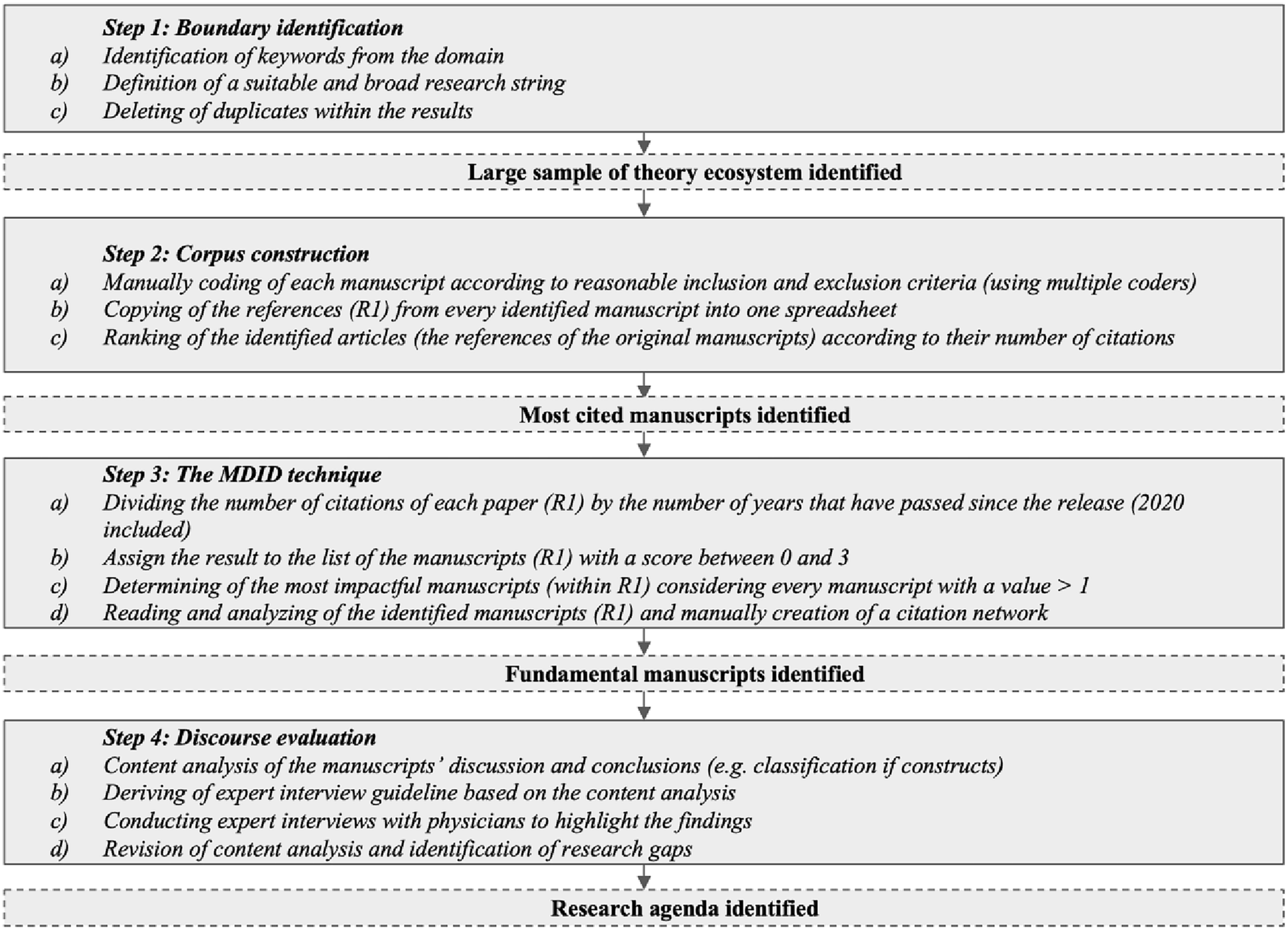
3.1 Modified discourse approach

For systematic literature analysis, new approaches are constantly being developed (vom Brocke et al. 2009, 2015). However, with the increasing number of publications, it is becoming more and more difficult to find a method that can provide a comprehensive picture of a discourse. The discourse approach is an instrument that creates a citation network based on fundamental manuscripts of a theory, a model, a framework, or a research domain (Larsen et al. 2019). It starts with the identification of fundamental theory-building papers (L1), followed by theory-contributing and other papers that cite these L1 papers (L2). In a last step, papers are identified by means of citations, which influenced the L2 papers (L3). Larsen et al. (2019) call the sum of these L1, L2, and L3 papers "the theory ecosystem."

However, it is not always obvious which manuscripts form the fundamental basis for a discourse. The discourse on the responsible use of AI in hospitals is a rather new one, as fundamental manuscripts have yet to emerge. Therefore, the discourse approach cannot always be applied exactly according to Larsen et al. (2019). We therefore propose a modified discourse approach. The aim of our approach is to start vice versa by identifying fundamental L1 manuscripts and to derive a research agenda for ethical considerations of AI in hospitals. As the IS perspective is rather interdisciplinary, we started our research to the field of IS and related disciplines using the litbaskets.io database with 3XL search. Our method consisted of four phases following the recommendations by Larsen et al. (2019) and highlighting the outcomes with interview findings. An overview of the applied research approach is provided in Fig. 1 and will be presented in the following sub-sections.

3.1.1 Boundary identification

A research domain is less a set of characteristics and more an evolving discourse between scholars (Larsen et al. 2019). To reflect this discourse, a starting point is first required. According to Larsen et al. (2019), this initial point is the origin of a theory, framework, or model. In this paper, however, we wanted to identify the status quo in research on ethical considerations on AI use in hospitals. Therefore, we based our boundary identification on elements of other systematic literature reviews such as a comprehensive keyword search as proposed by vom Brocke et al. (2015). To identify a first sample in the theory ecosystem, we first collected frequent keywords related to ethical principles of using AI in health and especially in hospitals. We selected artificial intelligence as the keywords as well as related terms that focus on more anthropomorphic forms of AI, because our focus was on technology that is also perceived as an AI by both the physicians and the patients. In addition, we selected ethic* and



Step 1: Boundary identification

a) *Identification of keywords from the domain*

b) *Definition of a suitable and broad research string*

c) *Deleting of duplicates within the results*

Large sample of theory ecosystem identified

Step 2: Corpus construction

a) *Manually coding of each manuscript according to reasonable inclusion and exclusion criteria (using multiple coders)*

b) *Copying of the references (R1) from every identified manuscript into one spreadsheet*

c) *Ranking of the identified articles (the references of the original manuscripts) according to their number of citations*

Most cited manuscripts identified

Step 3: The MDID technique

a) *Dividing the number of citations of each paper (R1) by the number of years that have passed since the release (2020 included)*

b) *Assign the result to the list of the manuscripts (R1) with a score between 0 and 3*

c) *Determining of the most impactful manuscripts (within R1) considering every manuscript with a value > 1*

d) *Reading and analyzing of the identified manuscripts (R1) and manually creation of a citation network*

Fundamental manuscripts identified

Step 4: Discourse evaluation

a) *Content analysis of the manuscripts' discussion and conclusions (e.g. classification if constructs)*

b) *Deriving of expert interview guideline based on the content analysis*

c) *Conducting expert interviews with physicians to highlight the findings*

d) *Revision of content analysis and identification of research gaps*

Research agenda identified

Fig. 1 Adapted discourse approach based on Larsen et al. (2019) to derive a research agenda

moral* as relevant keywords because they most precisely represented what we wanted to examine from a philosophical point of view. Furthermore, we selected common terms from the area of digital health. Afterwards we formulated a broad and comprehensive search string including the following terms:

> (AI or "artificial intelligence" or "chatbot*" or "chatbot*" or "conversational agent*" or "digital assistant*" or "virtual assistant*" or "personal assistant*" or "virtual agent*" or "ai-based system*") AND (health or "health care" or healthcare or "digital health" or "hospital*" or medicine or medical) AND ("ethic*" or "moral*")

We applied the search string on Scopus and used litbaskets.io (3XL search) to receive an interdisciplinary focused sample of manuscripts (Boell and Blair 2019). In addition, we manually searched for high-ranked conference articles (in International Conferences on Information Systems, European Conference on Information Systems, Hawaii International Conference on Systems Sciences, Americas Conference on Information Systems, Pacific Conference on Information Systems, Australasian Conference on Information Systems, and the German Wirtschaftsinformatik). In our initial sample, we focused on IS publications since our aim was to visualize and reflect the interdisciplinary discourse. However, as a basic search is not capable of providing a holistic overview, and we were also interested in retrieving literature outside the IS discipline, we conducted both a backward and forward search. In the backward search, we gathered the reference lists in the bibliographies of all the papers from the initial search and assessed their relevance regarding our research goal. Within the forward search, we considered every paper identified in the previous steps and analyzed literature that cited these identified papers after their initial publication. We thus expanded our search to other scientific domains and outlets. For example, we identified publications from healthcare (e.g., Journal of the American Medical Association) and philosophy (e.g., Philosophical Transactions of the Royal Society).

We conducted our literature search between September and October of 2020. After removing duplicates from the results, we identified 104 manuscripts as our initial sample. This sample consisted of interdisciplinary journals and high-level conference articles and was labeled as potential L2 articles (Larsen et al. 2019) who cite the fundamental manuscripts of the discourse on the ethical use of AI in healthcare.

3.1.2 Corpus construction

As a next step, we investigated the identified literature in more detail. Our aim was to understand the discourse on the ethical dimensions of AI in healthcare and especially in hospitals. We, therefore, manually scanned the 104 identified manuscripts according to their topic relevance. We excluded papers that did not directly address ethical dimensions and articles that did not address AI or AI-related technologies. We included manuscripts that covered both ethics and AI. Two experienced coders created a codebook and applied the exclusion and inclusion criteria to the manuscripts from the first search, following a title, abstract, and keyword scan method. This led us to 60 manuscripts that we considered the most relevant for the ethical discourse on AI in healthcare.

However, we knew that not all relevant articles for a discourse can be identified by a keyword search (Larsen et al. 2019). If a keyword search is too broad, it can lead to a list containing far more manuscripts than is practical to read; and if a keyword search is too narrow, that can result in missing highly relevant articles. To address these issues, we copied all references from these 60 manuscripts into one list, which led us to 2433 references. As our aim was to identify fundamental manuscripts for the ethical discourse on AI in healthcare, we ranked those references according to how often they were cited in the initially identified papers. The number of citations per paper within the list of all references is shown in Table 4 in the "Appendix".

3.1.3 Identification of fundamental manuscripts for the ethical discourse on AI in healthcare

Although the number of citations is an important indicator to measure the relevance of a manuscript within a discourse (Larsen et al. 2019), the time span between the publications also needs to be considered. To take publication time spans into account, we propose a manual detection of implicit domain (MDID) technique. We divided the number of citations of each paper by the number of years that have passed since the date of publication. This resulted in a score between 0 and 3 citations per year within the identified corpus. This score does not represent the overall citations per year of the manuscripts, but rather the number of times they were cited per year within the 60 papers that we identified as relevant for the ethical discourse on AI in healthcare. Among those, a few papers had a score > 1 and most of the papers scored lower than 1. The score describes the impact and relevance of the manuscript on the current discourse on AI in healthcare. To better understand the distribution of the scores, we visualized the dissemination of the scores in a graph. We found that there was a small group of manuscripts that stood out and scored higher than the majority of the articles. We identified these papers due to the visible threshold in the graph. This small group of papers scored 1.3 or higher and consisted of only 15 manuscripts. In addition, we manually scanned how these manuscripts were cited within the identified corpus of 60 papers to ensure that they were not only mentioned as a side note. We considered all of these 15 manuscripts as the fundamental articles. Additionally, these 15 manuscripts came closest to what Larsen et al. (2019) had described as L1 manuscripts. Those manuscripts are listed in Table 1.

As our aim was to understand and structure the ethical discourse on AI in hospitals, we further analyzed those manuscripts manually and created a citation network (Fig. 2). We scanned the manuscripts for common patterns and extracted the ethical principles for using AI in hospitals to provide a research agenda for academia.

3.2 Expert interviews

Besides using the discourse approach as a fruitful method to obtain a comprehensive picture of the knowledge within a certain domain (Larsen et al. 2019), we also conducted semi-structured expert interviews to highlight and underpin our findings. Expert interviews preserve knowledge from individuals with advanced experience in the research domain under investigation (Meuser and Nagel 2009). We thus initially defined criteria to find appropriate participants. Since discussions on

Table 1 Identified fundamental manuscripts of the discourse on the ethical use of AI in healthcare

Authors	Count	Score
Vayena et al. (2018)	8	2.667
Ting et al. (2017)	8	2
Char et al. (2018)	6	2
McKinney et al. (2020)	2	2
Zeng et al. (2019)	2	2
Yu et al. (2018)	5	1.667
Gulshan et al. (2016)	8	1.6
Reddy et al. (2019)	3	1.5
Yu and Kohane (2019)	3	1.5
Schiff and Borenstein (2016)	3	1.5
Parikh et al. (2019)	3	1.5
Luxton (2019)	3	1.5
He et al. (2019)	3	1.5
Froomkin et al. (2019)	3	1.5
Cath (2018)	4	1.334

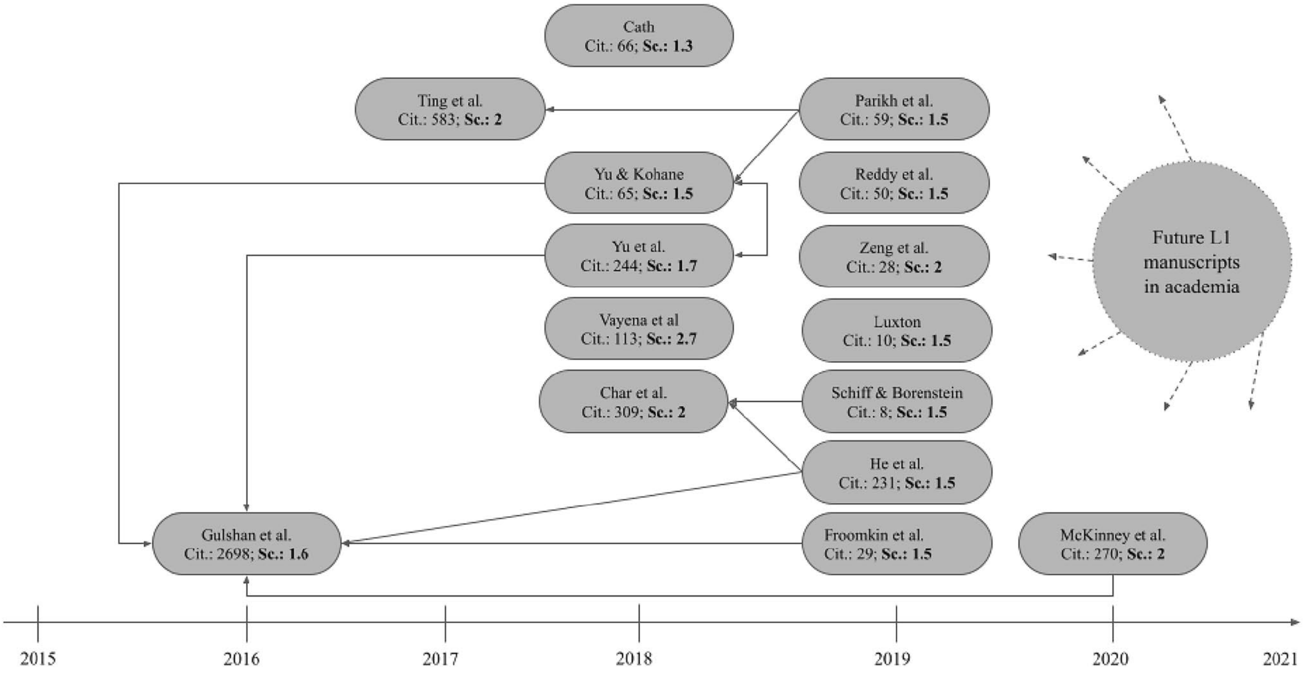


Fig. 2 Citation network of the 15 fundamental manuscripts

ethics in medicine are as ancient as the discipline itself, we intended to gain a holistic overview from experts of varying age groups. We further searched for medical experts working in hospital clinics who are frequently confronted with ethical questions impacting the well-being of patients. Following the recommendation of Creswell and Creswell (2018), three to ten individuals should be included for qualitative research. Moreover, we use the interviews to elaborate on and highlight our findings rather than to validate a theory. In total, we conducted six expert interviews with doctors and senior level experts in the context of hospital digitization from different medical disciplines. We interviewed one physician from obstetric care (resident doctor) and three surgeons from cranio-maxillofacial surgery (two senior physicians and one resident doctor). In addition, we spoke with a chief physician from a large hospital and a head of corporate communication with experience in digitization and change management in hospitals. An overview of our sample is outlined in Table 2. To guarantee anonymity of our interviewees, we used the synonyms E1–E6 in the following sections.

We used an open interview technique to provide the experts with enough room to elaborate on their subjective beliefs and experiences (Meuser and Nagel 2009). We structured the interview with a prefixed guideline (Table 6 in the "Appendix") with central questions referring to our research question (Qu and Dumay 2011). Initially, we described the interview process to the interviewee, including a short briefing of the study and the rights of the participants, followed

Table 2 Sample overview of expert interviews with physicians and senior level experts

Interviewee	Gender	Age	Tenure (years)	Position	Discipline	Hospital	Duration
E1	f	31	3.5	Resident doctor	Obstetric care	University Hospital of Frankfurt, Germany	28:17
E2	f	38	7	Senior physician	Cranio-maxillofacial surgery	University Hospital of Dusseldorf, Germany	31:38
E3	f	35	5	Senior physician	Cranio-maxillofacial surgery	University Hospital of Dusseldorf, Germany	30:30
E4	f	31	2	Resident doctor	Cranio-maxillofacial surgery	University Hospital of Dusseldorf, Germany	35:42
E5	m	67	20	Chief physician	Anesthesia	Retired	42:33
E6	m	44	17	Head of Corporate Communications	Digitization Think Tank	Clinical Center Dortmund, Germany	32:41

by a verbal consent to the interview being recorded. In the first official phase, we asked general questions on the expert's characteristics, current position, and duties within the practicing discipline. This helped us to understand the clinical environment of the expert while making the interviewee comfortable with the interview situation. The second phase served as a foundation to comprehend which ethical considerations physicians are confronted with and whether they follow a certain codex. Within the third phase, we asked question on what ethical problems technology in general might cause and how they are capable of resolving ethical issues. The fourth phase began by asking interviewees what they associate with AI. After receiving their answers, we provided a definition of AI to achieve the same level of knowledge among all participants for the remainder of the interview. We then asked specific questions about the application of AI in hospitals, e.g., how AI might support clinical processes, which factors are crucial for successful deployment, and which ethical guidelines AI must follow. In the fifth phase, the participants were asked to elaborate on future ways in which AI implementations in hospitals could improve the clinical procedures. The interview concluded by providing the interviewee with a chance to ask further questions or to provide additional information, followed by a debriefing by the interviewer.

The data were collected between September and October of 2020 by two researchers. As this period was still strongly influenced by the COVID-19 pandemic, all interviews were conducted via a virtual call. As we were not interested in the expert's substantive statements rather than physical gestures or facial expressions, we recorded the audio and not the video signal and, respecting data privacy protection, deleted the recordings once the analysis of the interview was finished. For the examination of the retrieved data, we conducted a qualitative assessment of content analysis as previously proposed (Schilling 2006). This helped us to reduce the volume of the data by removing unnecessary words to form short and concise sentences. We paraphrased the experts' explanations by carefully listening to each interview recording, then further generalized and reduced the contents, leading to comprehensive statements.

The analysis of the data was performed using a thematic analysis where paraphrasing was done shortly after the interviews were conducted. We derived deductive categories based on the constructs as identified from the discourse approach and used them as clusters (Glaser 2013). We thereby intended to obtain an understanding of the status quo and prospective orientations. This research approach can be classified as a descriptive-prescriptive procedure because experts described the situation, e.g., what has happened or what is happening now and what should happen in the future (Bear and Knobe 2016). Following the recommendations of (Gioia et al. 2013), we used short paragraphs or sentences as coding units, i.e. open coding. We used simple phrases or in vivo (second-order themes) to code the data, then categorized them under the constructs from the discourse approach (first-order theme). The coding process was collaboratively done by two researchers to distribute the effort of the analysis process, prevent a unilateral view of the data, and ensure intercoder reliability. Since the expert interviews were conducted with German participants working in German hospitals, the excerpts have been translated into English for the reader's understanding.

4 Results

We were able to identify 15 manuscripts that we could classify as fundamental by means of our modified discourse approach. The manuscripts were mostly published in medical journals, Nature, or Science (He et al. 2019; Parikh et al. 2019; Yu and Kohane 2019; McKinney et al. 2020). Among the papers, we found theoretical papers as well as empirical papers. Many manuscripts established principles for the ethical use of AI in hospitals or discussed different fields of application or types of AI. Although principles were strongly intertwined and we perceived some overlaps when directly comparing definitions between some papers, we could extract 18 unique ethical principles from the literature following Suddaby (2011). We consider these principles as mutually exclusive as they differed in their descriptions when comparing the 15 fundamental manuscripts. We classified the findings of our interviews into the four first-order themes beneficence, non-maleficence, justice, and autonomy and into the 18 s-order themes which represent the principles in Table 3.

One of the most mentioned ethical principles for using AI in healthcare was the principle of transparency (Cath 2018; Vayena et al. 2018; Zeng et al. 2019; Froomkin et al. 2019; He et al. 2019). It describes the visibility of the general logic of machine learning algorithms (Vayena et al. 2018). On the one hand, it is intertwined with the principles of explainability and explicability**,** which aim to not only make the algorithms transparent but also provide information for people with less technical knowledge such as patients or doctors (Cath 2018). Moreover, it seemed that explainability and transparency overlap and relate to similar issues. However, main difference between transparency and explainability is that transparency does not necessarily include further instructions such as a tutorial on how AI executes certain processes. If a hospital would provide access to the code of a system, they would provide transparency for this code; but to provide explainability, the code would need to be delivered with further explanation of its purpose and process. On the other hand, transparency is intertwined with the principle of fairness (Zeng et al. 2019). Zeng et al. (2019) stated that people in the context of healthcare might ask for transparency regarding the decision-making process of an AI out

Table 3 Ethical principles for the use of AI in hospitals extracted from the fundamental manuscripts

Type of issue	Principle	References	Description
Regulatory issues	Accountability	Cath 2018; Vayena et al. (2018), Zeng et al. (2019) and Reddy et al. (2019)	The determination of who is accountable for errors, who is socially responsible for the outcome of an AI, and which legal obligations have to be taken into account should be ensured
	Responsibility	Cath (2018), Char et al. (2018), Zeng et al. (2019) and Luxton (2019)	
	(Legal) liability	Schiff and Borenstein (2016), Vayena et al. (2018), Yu et al. (2018), Luxton (2019), and Reddy et al. (2019)	
	Privacy	Cath (2018), Vayena et al. (2018), Zeng et al. (2019) and He et al. (2019)	The protection of users' data and the compliance with general data protection regulations should be ensured
Normative issues	Avoiding bias and harms	Cath (2018), Char et al. (2018), Parikh et al. (2019), Reddy et al. (2019) and Yu and Kohane (2019)	The prevention of damage to one or more patients from the use of AI in healthcare should be ensured
	Patient safety	He et al. (2019) and Parikh et al. (2019) and McKinney et al. (2020)	
	Fairness	Cath (2018), Vayena et al. (2018) and Zeng et al. (2019)	The avoidance of discrimination of patients should be ensured using algorithmic fairness
	Informed consent	Schiff and Borenstein (2016), Ting et al. (2017) and Froomkin et al. (2019)	It should be ensured that physicians be able to explain the exact use of an AI to be sure that the patients know to what they are consenting
Technical issues	Interoperability and generalizability	He et al. (2019), Parikh et al. (2019) and McKinney et al. (2020)	It should be ensured that the training data for an AI represents a large population to provide interoperable and generalizable systems
	Iterative controllability and updatability	Yu and Kohane (2019)	It should be ensured that AI in hospitals is always controlled by trained physicians and updated with clinical workflow disruption
	Vigilance	Yu et al. (2018)	It should be ensured that responsible physicians frequently monitor the AI system
	Security	Zeng et al. (2019)	It should be ensured that the system has a certain level of robustness against cyber-attacks
Organizational issues	Feasibility and humanity	Gulshan et al. (2016), Char et al. (2018), Yu et al. (2018), Zeng et al. (2019) and McKinney et al. (2020)	It should be determined if and how AI is capable of improving care in hospitals
	Education of an AI-literate workforce	He et al. (2019)	It should be ensured that healthcare professionals are well trained and educated in the fields of medical informatics and statistics
	Interventions	Parikh et al. (2019)	It should be ensured that the output of a predictive AI is accompanied by guidance for medical interventions
	Explainability	Vayena et al. (2018), Yu et al. (2018) and Zeng et al. (2019)	It should be ensured that the use of AI in hospitals is understandable to the patient
	Transparency	Cath (2018), Vayena et al. (2018), Zeng et al. (2019), Froomkin et al. (2019) and He et al. 2019)	The visibility of the general logic of machine learning algorithms and its explanation should be ensured
	Trustworthiness	Yu and Kohane (2019)	It should be ensured that the patients and the physicians who use AI trust the systems' predictions

of concerns about fairness. However, we found no clear definition of what exactly fairness would mean in terms of AI and algorithms. We identified indications that in most fundamental manuscripts, the authors understand fairness as algorithmic fairness that ensures that there is no discrimination of minorities (Cath 2018). The results of the expert interviews confirmed the major relevance of transparency as an ethical principle. This especially refers to disclosing to medical experts how AI derives certain results. One expert clarified "I don't know if that is possible, but I should ideally understand what the AI is doing" (E4).

In addition to transparent communication about the presence of an AI, the liability must be clearly evident (Vayena et al. 2018). The principle of liability is closely linked to accountability and responsibility (Schiff and Borenstein 2016; Reddy et al. 2019). We summarized those three terms using responsibility as it was the most frequent and interchangeably used term within the considered literature. Accountability for errors that occur through AI use in hospitals has not yet been conclusively determined. One interviewee compared this to the debate on self-driving cars: "This reminds me of the debate about self-driving cars. It is unclear who is responsible. The car manufacturer? The insurance company? The software manufacturer? The driver? This has not yet been conclusively clarified with regard to AI in hospitals either" (E6). Liability can be defined as the legally obligated determination of who is morally responsible for medical errors regarding the use of AI (Schiff and Borenstein 2016). While liability tends to address the legal aspects, accountability is more focused on the authority to issue instructions. Responsibility, on the other hand, includes an ethical and social component and addresses the questions of how much indirect responsibility is relevant and which actors are indirectly responsible. However, liability, responsibility, and accountability are not clearly delineated in most of the fundamental works and need further definitions, clarifications, and delimitations (Reddy et al. 2019). While the terms are often used synonymously, they can also sometimes be used too narrowly. In a case study, Luxton (2019) examined the ethical, responsible, and legal liability issues surrounding the use of IBM Watson in hospitals. They provided a guide for physicians who want to use AI tools in hospitals and identified precautions based on a case where patients with leukemia should be treated. The interviews revealed that while AI can be helpful in making suggestions, medical experts should be responsible for health-related decisions. One expert summarized "the human emotional aspects are simply missing. AI simply cannot consider every human aspect" (E3). Another expert added that "physicians possess numerous years of experience. Subjective human impressions might positively influence the treatment. There is still quite some information that an AI does not or cannot have." (E6). Mentioned examples included the family background or health insurance.

Another reason why transparency regarding how algorithms work is highly ethically relevant is that the training dataset of an AI can influence the system's output (Parikh et al. 2019). That means that algorithms trained on a specific group of patients (e.g., in a specific clinic of one city) may not be generalizable and interoperable. Therefore, when using AI in hospitals, generalizability should be ensured (He et al. 2019; McKinney et al. 2020) to avoid unintended outcomes that could potentially harm patients' health. If an AI is too specialized on one task in one environment, it could deliver wrong treatment assistance when being transferred to another context. Generalizability could in this case be ensured if an AI would be tested in a multiple-case study.

When using AI in healthcare, most authors mentioned the avoidance of bias and harms as an important principle for physicians (Cath 2018; Char et al. 2018; Parikh et al. 2019; Reddy et al. 2019; Yu and Kohane 2019). Schiff and Borenstein (2016) discussed potential harms emerging from interactions between humans and AI when AI is considered as part of a medical team. They specifically discussed how responsibility should be distributed among physicians, developers, and further stakeholders, and they further provided advice for practitioners. Overall, we did not find much information or guidance on what exactly is possible harm and which precautions could be taken to avoid harm to patients. What we found was that education of an AI-literate workforce would play an important role when deploying an AI in a clinical environment (He et al. 2019). The introduction of an AI should therefore always involve all affected stakeholders, and all junior physicians need to be trained and educated in the areas of medical computer science and statistics (He et al. 2019). One expert explained, "I think especially young or unexperienced doctors benefit or learn from AI-based decisions. Experienced physicians have the most important parameters for the evaluation of certain disease in their heads, but this does not apply to novice physicians" (E6). In addition, the output of a predictive AI system in a health context should provide guidance for concrete medical interventions to explain the output of the prediction to physicians (Parikh et al. 2019).

One specific type of harm that was discussed in the fundamental articles was potential privacy issues (Cath 2018; Vayena et al. 2018; Zeng et al. 2019; He et al. 2019). However, we neither found detailed information on what exactly are the relevant privacy issues regarding AI use in healthcare, nor information on how possible issues could be addressed. One example could be an AI asking for sensible information that patients do not want to reveal.

When patients need to consent to the use of AI for a treatment or a therapy, they need to have trust in the system and the controlling physicians (Yu and Kohane 2019). Trust could be achieved through a high level of transparency and explainability. One important principle, related to transparency and explainability, is the informed consent process (Schiff and Borenstein 2016; Ting et al. 2017; Froomkin et al. 2019). To be able to agree to informed consent, the patient must understand how an AI is used and what consequences the use of an AI might have (e.g., on a treatment). Patients thus must to be made aware of the fact that some kind of AI is involved in their treatment or course of disease. One expert testified "In principle, the patient must agree to be 'treated' by an AI. This also implies explaining what this technology is doing and related consequences" (E6). This can be complicated for several reasons (Schiff and Borenstein 2016). First, the physician must have sufficient knowledge to explain the use of AI. Second, it is often difficult even for experts to understand the exact procedure of AI (black-box problem), since very large amounts of data and computing capacity are involved. One expert highlighted "We already heavily rely on certain technology. AI might yield in thinking less thus being less involved and losing the feeling of being responsible" (E2). Strategies to counteract this process could not be found in literature and need to be further investigated.

Yu and Kohane (2019) argued that the data and the algorithms need to be frequently controlled and updated to address the clinical workflow disruption. This requires not only the possibility of checking and updating, but also a continuous vigilance by the responsible physicians in hospitals (Yu et al. 2018). Not only does the system need to be checked and updated, but the feasibility of using AI in hospitals should be regularly updated as well (Gulshan et al. 2016; Yu et al. 2018; McKinney et al. 2020). It should be determined how exactly the use of AI would lead to an improvement in care (Gulshan et al. 2016). If the system is determined feasible and beneficial, the AI also needs to be checked for security issues to avoid cyber-attacks and errors (Zeng et al. 2019). Cyber-attacks could result in privacy violations, data misuse and even physical harm of patients through data and system manipulations.

We provide an overview of the ethical principles we extracted from the 15 fundamental manuscripts in Table 3. As not all principles were described in detail, we added some aspects of our understanding in the descriptions. Some principles were used interchangeably, which is why we provided just one description for up to three principles in some cases. We categorized the principles according to the types of issues that they may address. By regulatory issues, we refer to ethical issues that require clear rules and possible legal guidance, such as determining who is responsible for errors made by AI-assisted treatment. Normative issues are those that cannot be clearly defined by rules and laws, but should be guided by social norms (e.g., which patients should be treated first). As technical issues, we consider all types of issues that are caused by design (mostly unintentionally), such as a biased training dataset. Organizational issues are problems that could be addressed by restructuring processes within a hospital such as a lack of technical expertise of physicians, which could result in not being able to explain an AI-based treatment assistant.

In addition to the relationships between the ethical principles of AI discussed within the 15 fundamental manuscripts, we identified the citation structures between the articles. We found that the citations within our identified discourse ecosystem often differed from the citations of an article on Google Scholar or meta-databases such as Scopus meaning that the most cited manuscripts on these databased were not the ones that centrally discussed on ethical issues of using AI in hospitals. This highlights the importance of this modified discourse approach. The time span of the manuscripts we considered relevant for the ethical discourse on AI in hospitals ranged from 2016 to 2020. Most articles we identified were published in 2019. Ten of the articles formed a citation network, whereas five of the articles did not cite or were not cited by any of the other manuscripts. The most cited article within the identified network was also the most cited article on Google Scholar and Scopus on the topic of ethical frameworks of AI within healthcare. The most cited article within our identified papers was an empirical work and did not focus on theorizing on ethics and AI in hospitals (Gulshan et al. 2016). However, its findings, mentioned limitations, and conclusions were often used as a starting point for ethical discussions. In Fig. 2, we provide a timely overview of how the fundamental manuscripts cited each other and visualize ways in which future research could contribute to this network by referring to these valuable articles and connecting them to a holistic picture. For each fundamental article, we present the Google Scholar citations and the score in our network. The arrows symbolize a citation within the network and the dotted arrows offer possible points of reference for future research. Although some of the manuscripts cited each other, we found no article that discussed the others in light of ethical challenges and problems in hospitals. Rather, the articles often used different terms to describe similar aspects without referring to each other and did not specify important aspects.

5 Discussion

Applying the modified discourse approach proposed by Larsen et al. (2019), we identified 15 manuscripts that are fundamental for the discourse on the ethical dimensions of using AI in hospitals. Although AI and healthcare are important application fields in many disciplines, we did not find one discipline that clearly stood out. Furthermore, the identified manuscripts made little reference to each other (see Fig. 2). Although we found papers such as Gulshan et al. (2016), which were cited more frequently among the fundamental manuscripts, these were empirical papers rather than contributions to the ethical discourse in the use of AI in hospitals. However, in our identified network, we could not detect any established work reflecting the current discourse in academia or considering the opinions of physicians with regard to ethical considerations and dimensions of AI. With this work, we address this issue (RQ 1). In addition, we provide a research agenda in the next chapter that aims to guide academia in future works (RQ 2).

We also found that the discourse did not followed a logical structure. Five articles we considered did not refer to any other manuscripts that we classified as fundamental (Cath 2018; Vayena et al. 2018; Zeng et al. 2019; Luxton 2019; Reddy et al. 2019). This could lead to parallel discussion streams on the same topic. Interestingly, the most cited manuscript among the fundamental manuscripts was an empirical work that addressed ethical dimensions in a limited way and only within the conclusion and limitations (Gulshan et al. 2016).

Most identified articles either provided an incomplete view of the ethical challenges of applying AI in hospitals or functioned as empirical works that just scratched the surface of ethical principles and issues. Some of the existing articles focused on ethical challenges of very narrow AI technologies and did not consider a bigger picture (Gulshan et al. 2016; Ting et al. 2017; McKinney et al. 2020). On the other hand, some of the articles tried to derive ethical principles for the use of AI in healthcare which did not really differ from general ethical principles for using AI (Cath 2018; Vayena et al. 2018; Zeng et al. 2019).

Considering the fundamental manuscripts, no article focused on an overarching moral principle such as virtue ethics. Rather, the ethical perspective was not clearly defined. In the context of the ethical use of AI in hospitals, this could be deeply problematic, as virtues can be used to provide guidance to an AI-based system about what is right and wrong (Siau and Wang 2020). Future research needs to build on ethical perspectives similar to how moral virtues are discussed by Beauchamp and Childress (2019) and transfer these considerations to the context of AI applications in hospitals. Our research aims to guide this process.

Most of the principles we found were not discussed in detail and did not address the actual use of AI in hospitals (Char et al. 2018). In many articles, the same aspect was discussed using different terms such as explicability and explainability (Floridi et al. 2018; Vayena et al. 2018; Yu et al. 2018; Zeng et al. 2019) or accountability (Cath 2018; Vayena et al. 2018; Zeng et al. 2019; Reddy et al. 2019), responsibility (Cath 2018; Char et al. 2018; Zeng et al. 2019; Luxton 2019) and liability (Schiff and Borenstein 2016; Vayena et al. 2018; Yu et al. 2018; Luxton 2019; Reddy et al. 2019). In addition, ethics principles for using AI in healthcare are often intertwined and cannot be considered separately. However, we hardly found any discussion regarding dependencies between principles. Furthermore, detailed explanations on how ethical principles can be defined in the context of AI in hospitals were limited. Most principles lacked further definitions or were described on a meta-level that did not take into account ways in which they could be applied in healthcare. We, therefore, provide knowledge on how the principles should be examined and extended in future research. In Fig. 3, we show a structure that is more applicable for further research with dependencies of different levels of ethical principles for the use of AI in hospitals. Based on the relationships between ethical principles in the context of AI in hospitals, we provide a research agenda for academia.

6 A research agenda for academia

A philosophical perspective that specifically addresses ethical dimensions of AI in hospitals does not appear in the current discourse; although it cannot be dismissed that individual papers exist that address this topic. Researchers from various disciplines need to include this ethical perspective in their future work, as philosophical venues are classically the drivers of ethical discussions. Within the identified manuscripts, we found different categorizations of ethical principles for AI. For ethical dimensions of using AI in hospitals, however, we could not find a common understanding of how to structure ethical principles. Therefore, we propose a research agenda for academia whose structure is based on the widely known articles from Beauchamp and Childress (2019) on biomedical ethics and Floridi et al. (2018), who applied these principles to provide an ethical framework for a moral AI society. We argue that although the same categories of biomedical ethics are relevant for considering ethical dimensions of using AI in hospitals, their definition and compliance are not clearly actionable in further research nor in medical practice. As an overarching moral principle, we focused on a virtue ethics perspective as suggested by Siau and Wang (2020).

With our research agenda, highlighted with the results from the expert interviews, we aim to guide future research to ensure that researchers theorize and discuss the most important issues and challenges of using AI in hospitals. With their knowledge, interdisciplinary scholars will be able to provide guidance for physicians who must make the decisions about the use of AI in hospitals. On the other hand, they can also ensure that AI is used by hospitals for the benefit of patients and not in the interests of, for example, hospital profitability. Based on the suggestions of Beauchamp and Childress (Beauchamp and Childress 2019), we structured our research agenda into the categories of beneficence, non-maleficence, justice, and autonomy. Future research can address either one of these categories or one of the four issue types from Table 3. For more applied work, we recommend addressing the issue types; for theoretical and philosophical work, we recommend addressing the categories of bioethical principles.

To provide guidance for future research, we propose the following research questions (Table 4), which are structured according to the four bioethical principles (Beauchamp and Childress 2019).

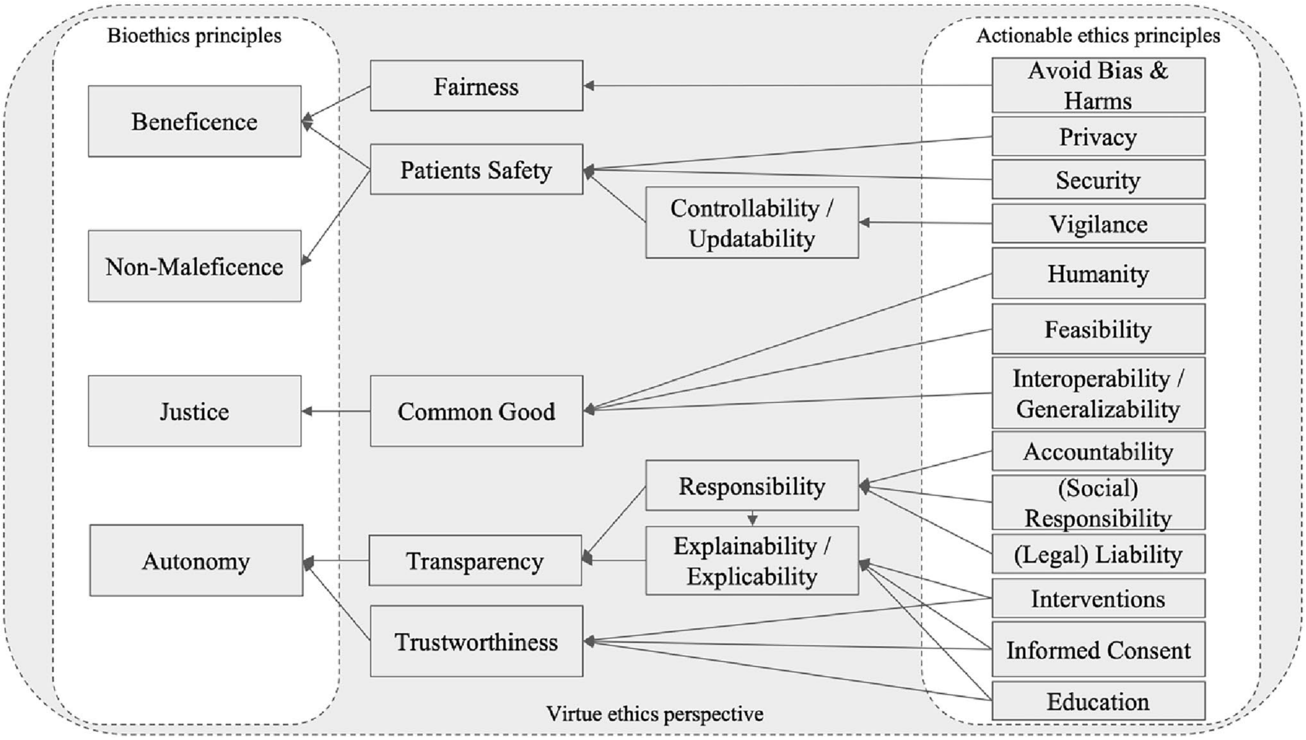


Fig. 3 Visualization of the relationship between actionable ethical principles for using AI in hospitals and bioethical principles according to Beauchamp and Childress (2019) and Floridi et al. (2018)

6.1 Beneficence

Floridi et al. (2018) defined beneficence as a principle that ensures that an AI promotes the well-being of humans and its output favors the common good. But what does this mean in the context of using AI in hospitals? While AI should act in a fair way (Cath 2018; Vayena et al. 2018; Zeng et al. 2019), it is not clear exactly what this implies. Further research should address in more detail the aspect of fairness in the field of AI implementation in hospitals. This would ensure the beneficence of the system in favor of the patients. Fairness can be achieved by avoiding bias and harm to all patients. For example, the use of AI should not exclude certain minority groups (e.g., people with rare diseases). One expert emphasized, “There are also ethical differences within cultures. In some countries, abortion is simply not an option for women” (E1). Previous research has highlighted cases where AI delivered poor predictions in healthcare due to biased data (Vayena et al. 2018). There are data sources that do not represent the true epidemiology within a given demographic, for example in population data biased by the entrenched overdiagnosis of schizophrenia in African Americans. In this cases AI needs mechanisms to detect incomplete or biased data. However, research on this is rare. Although some studies have detected unfair behavior of AI in hospitals, limited research has been conducted on the prevention of such issues. Using rich dataset training data for an AI could be one approach to avoid unfairness in hospitals; but how this can be achieved is a question that should be addressed. The same applies for AI violating patients’ safety. Previous research has stated that patients’ safety is an important factor for deciding whether an AI-based system can be used or not (Char et al. 2018; Zeng et al. 2019; He et al. 2019; Parikh et al. 2019) and discussed cases where it was violated. However, research on how to ensure patients’ safety when subjected to AI treatment assistants is still rare. One expert underlined that, “AI should support with difficult therapy decisions securing the well-being of patients, for example, whether palliative or radiation treatment is more appropriate” (E3).

6.2 Non-maleficence

AI use in hospitals should also be non-maleficent (Floridi et al. 2018). In contrast to beneficence, which includes what an AI should do, the principle of non-maleficence aims to avoid ethical issues when using AI e.g., in hospitals. However, in previous research, we did not find a comprehensive picture of the spectrum of possible maleficence caused by AI in hospitals. Due to the black-box character of AI, it is almost impossible to predict all consequences of its use, but the current state of knowledge could be depicted. It also remains unclear how non-maleficence in hospitals can be ensured when using AI. We could derive the following aspects from

Table 4 Formal grouping of research questions to guide future research on ethical dimensions of AI in hospitals

Bioethical principles	Actionable principles	Exemplary research questions
Beneficence	Vigilance Security Privacy Avoid bias and harms	1. How can the principle of fairness be defined in the context of using AI in hospitals? 2. Which medical data should be used to derive AI recommendations for therapeutic and treatment processes? 3. How can AI systems inform decisions made by healthcare professionals? 4. How can disadvantages to patients belonging to certain minority groups be removed or reduced? 5. In which application domains of digital health can AI be introduced as decision support systems to enhance hospital procedures and patient treatment? 6. To what extent can AI assist with difficult therapy decisions for certain patient groups?
Non-maleficence	Privacy Security Vigilance	1. What are possible harms caused using AI in hospitals? 2. How can bias within the medical data used by AI be recognized and resolved by healthcare professionals? 3. How could a control mechanism for decision support for physicians through AI in hospitals be designed and developed? 4. How can the awareness of vigilance regarding AI used in hospitals be increased? 5. How can it be ensured that medical information is not retrieved by third parties? 6. To what extent can external data manipulations within AI datasets be detected and prevented by physicians?
Justice	Humanity Feasibility Interoperability/generalizability	1. How can AI applications in hospitals contribute to the common good of a society? 2. How can common good be defined and interpreted by AI applied in clinical environments? 3. Which guidelines are essential to ensure common good when using AI in hospitals? 4. To what extent can physicians be psychologically relieved of moral dilemmas when using AI in hospitals? 5. How is AI able to improve the doctor-patient relationship in hospitals? 6. How can existing AI applications in hospitals be transferred to other conditions, departments, countries, and cultures? 7. To what extent are generalizable AI results ensured?
Autonomy	Accountability (Social) Responsibility (Legal) Liability Interventions Informed consent Education	1. To what extent do physicians perceive themselves to be losing their autonomy when AI is applied in hospitals? 2. How should the application of AI in hospitals be transparently presented to medical experts and patients? 3. Who can be held accountable and socially responsible for AI-driven decisions, and under which clinical conditions? 4. How can the legal liability for using AI in hospitals be clarified and implemented in a legal foundation? 5. Who is accountable and responsible for ensuring legal alignment when using AI in hospitals? 6. How can AI accompany its outputs with concrete recommendations for use in medical interventions? 7. How can it be ensured that both the physicians and the patients are aware of the consequences when consenting to the use of AI in a hospital? 8. How should AI applications be designed to be utilized only under voluntary conditions among clinicians and patients? 9. How do we need to educate and train physicians to ensure an ethical use of AI in hospitals? 10. What kind of training increases trustworthiness in using AI in hospitals?

the literature that refer to non-maleficence: patients' safety, privacy, security, controllability, updatability, and vigilance (Cath 2018; Char et al. 2018; Vayena et al. 2018; Yu et al. 2018; Zeng et al. 2019; He et al. 2019; Parikh et al. 2019; Yu and Kohane 2019; McKinney et al. 2020).

When applying AI in a hospital, possible violations of patients' privacy must be identified and solutions need to be developed. However, AI could also cause physical damage to patients' health, for example, when delivering decision support for diagnoses or medications. Although the highlighted training dataset is also potentially relevant for this, future research needs to determine which decisions could be supported by AI and how this decision support could be controlled. However, it seems that the decision support, e.g., regarding treatment recommendations, should always be monitored and assessed by human physicians: "It will never be the case that an AI takes over the complete diagnosis. It will always be the case that there is a choice and the human being decides at the end of the day" (E6). The technical controllability and updatability of a system, as well as the vigilance of the physicians, need to be ensured. In addition to monitoring AI for internal errors, we identified ethical issues regarding the external security of a system. For example, cyber-attacks could manipulate the data basis of an AI without the users noticing. Therefore, future research needs to address these types of security issues when using AI in hospitals. This leads us to the following further research questions: How can awareness for vigilance be increased?

6.3 Justice

The principle of justice covers aspects that "contribute to global justice and equal access to the benefits" for individuals and society (Floridi et al. 2018). In the literature, we found overlaps with the principle of fairness that aimed at avoiding any type of discrimination (Cath 2018; Vayena et al. 2018; Zeng et al. 2019). For a sharper demarcation, however, in this article, we focus on the aspect of common good when mentioning fairness. Future research should investigate what common good exactly implies and how common good can be achieved by AI. This might contain "psychological relief from doctors in the context of a triage" (E2), i.e., classification of patients in a crisis according to the severity of the injuries, but also "improving the doctor-patient relationship when AI handles standard procedures" (E4). In the literature of fundamental manuscripts on the ethical dimensions of AI in hospitals, we found four actionable principles that can be assigned to common good and justice: humanity, feasibility, interoperability, and generalizability (Gulshan et al. 2016; Char et al. 2018; Yu et al. 2018; Zeng et al. 2019; He et al. 2019; Parikh et al. 2019; McKinney et al. 2020). Future research should investigate which AI applications in hospitals can benefit humanity. Furthermore, for each AI application, the technical feasibility of the application for the common good needs to be evaluated. In many cases, AI technologies in hospitals are only used for a very specific case within a system, e.g., in angiography: "There are AI-based systems, for example in angiography, which determine with a certain probability and based on certain points that are detected within a vessel, what the rest of the vessel might look like" (E6). Future research should focus on how to make these AI systems interoperable and how to make the outputs of an AI-based system in hospitals more generalizable.

6.4 Autonomy

As another principle of bioethics, autonomy is defined as the right of patients to make decisions about their treatments, which implies that they mentally understand the situation (Beauchamp and Childress 2019). With AI, the question arises how patients' autonomy can be ensured as we willingly "cede some of our decision-making power to machines" (Floridi et al. 2018, p. 698). Future research should focus on how autonomy has to be ensured when using AI as support for a treatment and how this autonomy can be achieved. One expert explained, "the patient is in the center of attention" (E1) and further "as a physician you cannot evade responsibility" (E4).

In the literature, we found two fundamental principles by which autonomy can be achieved: transparency (Cath 2018; Vayena et al. 2018; Zeng et al. 2019; Froomkin et al. 2019; He et al. 2019) and trustworthiness (Yu and Kohane 2019). If patients have transparency about the use and application of AI in a hospital on the one hand, and trust in the way it works on the other hand, autonomy can be achieved. One way of achieving trust is to "show the power behind it. If you do studies like the one with Watson and show comparatively that an AI achieves several times better results than a human expert, then that naturally creates trust" (E6). According to E6, presenting the advantages of accompanied studies could be an adequate strategy to increase trustworthiness. However, to ensure adherence to both principles, more detailed aspects must be considered. Transparency does not only imply that a patient is informed about whether AI is being used and could potentially understand how it works. Transparency also includes explaining to the patient exactly how an AI-based system works and how its use might affect his or her treatment (Vayena et al. 2018; Yu et al. 2018; Zeng et al. 2019). This requires considering not only the principle of explicability, but also the principle of responsibility. The patients must be aware of who is responsible for the consequences and outputs of the use of AI in a hospital. We found three types of responsibility that future research should examine more closely: functional accountability (Cath 2018; Vayena et al. 2018; Zeng et al. 2019; Reddy et al. 2019), social responsibility (Cath 2018; Char et al. 2018; Zeng et al. 2019; Luxton 2019), and legal liability (Schiff and Borenstein 2016; Vayena et al. 2018; Yu et al. 2018; Luxton 2019; Reddy et al. 2019). This is also in accordance with E6, who stated, "The question of responsibility has not yet been conclusively clarified and is, therefore, philosophical to a certain degree. We as company are accountable for keeping our stable clean. But we should also have the doctors who can also question this again in case of doubt. But a certain amount of legal liability should also lie with the manufacturer, who should also be responsible for ensuring that the AI is always up to date."

Future research should, therefore, look at who is operationally responsible for AI and who has the authority to issue instructions on the use of AI, as well as who may not be directly responsible for the consequences of the use of AI but should be involved from an ethical perspective. In addition, it should be further investigated how the legal framework for the use of AI in hospitals should be designed and how it can be ensured that both physicians and patients are aware of it. A precise explanation of responsibility is part of the explainability of the ethical framework. How exactly this explainability can be ensured has not yet been sufficiently researched. We found three actionable principles that could enable explainability of AI: interventions (Parikh et al. 2019), informed consent (Schiff and Borenstein 2016; Ting et al. 2017; Froomkin et al. 2019), and education (He et al. 2019). Future research should address the fact that the use of AI should always be accompanied by concrete recommendations for interventions by physicians, as they must interpret the AI's outputs. Further research is also needed to determine exactly how these interventions should be designed. Another sub-area of ethical research in AI is informed consent. Future research should explore ways to ensure that physicians explain the effects of the use of AI to patients well enough to enable confident decisions on whether to consent or refuse. However, to ensure explicability, physicians need to be trained in this matter. Future research should explore in more detail what types of training and education are needed to enable the explainability of AI to the patient. Interventions, informed consent, and education are also important components in creating trustworthiness. Future research should explore how exactly trust can be created in a system on the part of physicians and patients. However, trust in AI must be treated with caution, as clinicians "rely on the technology and become dependent on it" and further, "AI does the thinking and people act blindly" (E2).

7 Conclusion and limitations

In this article, we presented the current discourse in the domain ecosystem of ethical considerations on AI in hospitals. Drawing from theoretical foundations (i.e., Beauchamp and Childress 2019; Floridi et al. 2018), enlightened by semi-structured expert interviews with clinicians, this article contributes to theoretical foundations by presenting research areas that need to be faced when AI is used in hospitals. These results are highly relevant for practitioners, academia, and healthcare researchers and inform societal issues and challenges.

The main theoretical contribution of this research is the proposal of a research agenda explaining where in-depth investigations are needed. Our study demonstrates that current research scratches the surface rather than conducting profound examinations. We thus guide scholars' efforts for future studies and encourage the prospective discourse of ethical considerations of AI in healthcare. On a practical level, physicians comprehend to what extent the application of AI in hospitals seems fruitful as well as where ethical questions arise that could affect patients' physical and psychological well-being. We, therefore, aim to raise practitioners' awareness for the possible up- and the downsides of AI in healthcare. In terms of implications for society, individuals realize that ethical considerations of AI are vital, as the overall well-being of patients has the highest priority among clinicians.

As with all research, certain limitations apply. Since we aimed to identify highly relevant and fundamental theory-building papers (L1), we did not take a closer look at other papers citing these publications (L2, L3). In total, we have identified 15 fundamental articles, providing a sufficient foundation for our research agenda. However, it is possible that we could have missed some relevant literature investigating ethical considerations and dimensions of AI in hospitals, which may have provided additional knowledge. Moreover, we retrieved articles from interdisciplinary outlets and conducted a forward as well as backward search to obtain relevant publications from related disciplines. Even though the fundamental theory-building papers are from various disciplines and thus provide transferable results, publications from other sources (i.e., PubMed, an essential database for biomedical literature) might have yielded additional insights. Furthermore, the group of experts we interviewed was quite homogenous, with a small number of individuals that only cover a limited fraction of knowledge. Interviewing additional hospital employees, i.e., clinicians from other departments or employees working in other hierarchies as nursing staff, might have led to a more holistic picture.

We invite scholars to address the exemplary research questions we have provided in this article in the context of the bioethical principles. The citation network of the 15 fundamental manuscripts can be used as a starting point to better highlight the ethical discourse of AI in hospitals and to extend and deepen our discussion. We suggest that researchers consider virtue ethics as the main ethical perspective, as virtues need to be defined when AI-based systems are applied for treatment support in hospitals. The 18 ethical principles we found, and especially the 13 actionable principles, contribute to the discourse of AI use in hospitals and can serve as guidance for academia as well as physicians and healthcare decision-makers.

Appendix

See Tables 5 and 6.

Table 5 Ranking of identified articles according to their number of citations

Number of citations	Number of papers
8	2
6	2
5	2
4	7
3	23
2	115
1	2713

Table 6 Interview guideline (German interview questions have been translated into English)

Phase	Research goal	Questions
Briefing	Welcoming the interviewee and providing general information about the research and brief introduction to the topic	–
Demographic data	Getting an understanding of the interviewee including position within the hospital and the areas of responsibility	a. Could you please introduce yourself? b. What is your current position in the hospital? c. What responsibilities does your position involve? d. How long have you been working in this position / in this hospital?
Ethical considerations in healthcare and hospitals	Ethical considerations physicians are confronted with and whether they follow a certain codex	a. What ethical considerations are you confronted with in your everyday work? b. What is the ethical code you follow?
Ethical considerations and technology	Ethical problems technology raises and how they are capable to resolve ethical issues	a. Which technologies are used in your hospital to support your work? b. Which technologies do you rely on for your decisions? c. Which ethical problems can technology cause? What questions arise? d. Which ethical problems can a technology help to solve?
Ethical considerations and AI	Specific questions on the application of AI in hospitals and which factors are crucial for a deployment and what ethical guidelines must be follow	a. What do you associate with the term "artificial intelligence"? Providing an explanation of AI and current examples to assume the same knowledge among all participants b. For which tasks can AI be used as support in hospitals? c. Which tasks can AI be allowed to take over independently and which not? d. Which factors must AI consider when being used hospitals? Which rules must be obeyed? e. What is AI not allowed to decide for itself? What outcomes need to be prevented? What negative consequences may result? f. What are ethical conditions, requirements, and challenges for the application of AI in hospitals? g. Which morally reprehensible decisions should AI not derive? h. Which moral decisions could an AI make better compared to a human being?
AI and future perspectives	Future ways of AI implementations in hospitals improving clinical procedures	a. For what purposes would you use like to use AI in hospitals? b. Which decision would you rather follow, that of a human or an AI? Please elaborate c. How do you think is the role of AI in hospitals changing in the future?
Debriefing	Debriefing of the interviewee and explanation of the research background, possibility for the interviewee to ask further question or giving closing remarks	a. What other question did you expect but was not asked? b. Do you have further questions / comments on the topic?

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00146-021-01239-4.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

AEM (2020) Akademie für Ethik in der Medizin—Ziele und Aufgaben. https://www.aem-online.de/. Accessed 29 Apr 2021

Alami H, Lehoux P, Auclair Y et al (2020) Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 22:e17707. https://doi.org/10.2196/17707

AMA (2020) American Medical Association Principles of Medical Ethics. https://www.ama-assn.org/delivering-care/ethics/code-medical-ethics-overview. Accessed 29 Apr 2021

Amnesty International (2020) Help women and girls in Poland fight dangerous new restrictions on abortion. https://www.amnesty.org/en/get-involved/take-action/help-women-and-girls-in-poland-fight-new-restrictions-on-abortion/. Accessed 15 Apr 2021

Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28:15–26

Arnold T, Scheutz M (2018) The "big red button" is too late: an alternative model for the ethical evaluation of AI systems. Ethics Inf Technol 20:59–69. https://doi.org/10.1007/s10676-018-9447-7

Atherton PJ, Smith T, Singh JA et al (2013) The relation between cancer patient treatment decision-making roles and quality of life. Cancer 119:2342–2349. https://doi.org/10.1002/cncr.28046

Bærøe K, Miyata-Sturm A, Henden E (2020) How to achieve trustworthy artificial intelligence for health. Bull World Health Organ 98:257–262. https://doi.org/10.2471/BLT.19.237289

Bargshady G, Zhou X, Deo RC et al (2020) Enhanced deep learning algorithm development to detect pain intensity from facial expression images. Expert Syst Appl 149:113305. https://doi.org/10.1016/j.eswa.2020.113305

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J et al (2020) Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Batin M, Turchin A, Markov S et al (2017) Artificial intelligence in life extension: from deep learning to superintelligence. Informatica 41:401–417

Bear A, Knobe J (2016) Normality: part descriptive, part prescriptive. Cognition 167:25–37. https://doi.org/10.1016/j.cognition.2016.10.024

Beauchamp TL, Childress JF (2019) Principles of biomedical ethics, 8th edn. Oxford University Press, New York

Bell DE (1989) Decision making: descriptive, normative, and prescriptive interactions. Cambridge University Press

Bickmore T, Puskar K, Schlenk E et al (2010) Maintaining reality: relational agents for antipsychotic medication adherence. Interact Comput 22:276–288. https://doi.org/10.1016/j.intcom.2010.02.001

Boell SK, Blair W (2019) An IT artifact supporting exploratory literature searches. In: Australasian conference on information systems. http://www.litbaskets.io. Accessed 21 Jun 2021

Bore M, Munro D, Kerridge I, Powis D (2005) Selection of medical students according to their moral orientation. Med Educ 39:266–275. https://doi.org/10.1111/j.1365-2929.2005.02088.x

Brachten F, Brünker F, Frick NRJ et al (2020) On the ability of virtual agents to decrease cognitive load: an experimental study. Inf Syst E-Bus Manag 18:187–207. https://doi.org/10.1007/s10257-020-00471-7

Brendel AB, Mirbabaie M, Lembcke TB, Hofeditz L (2021) Ethical management of artificial intelligence. Sustainability 13:1–18. https://doi.org/10.3390/su13041974

Burton RJ, Albur M, Eberl M, Cuff SM (2019) Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. BMC Med Inform Decis Mak 19:1–11. https://doi.org/10.1186/s12911-019-0878-9

Bygstad B, Øvrelid E, Lie T, Bergquist M (2020) Developing and organizing an analytics capability for patient flow in a general hospital. Inf Syst Front 22:353–364. https://doi.org/10.1007/s10796-019-09920-2

Cath C (2018) Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philos Trans R Soc A Math Phys Eng Sci. https://doi.org/10.1098/rsta.2018.0080

CEOM (2020) Principles of European Medical Ethics. http://www.ceom-ecmo.eu/en/view/principles-of-european-medical-ethics. Accessed 29 Apr 2021

Char DS, Shah NH, Magnus D (2018) Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 378:979–981

Crawford K, Calo R (2016) There is a blind spot in AI research. Nature 538:311–313

Creswell JW, Creswell DJ (2018) Research design: qualitative, quantitative, and mixed methods. SAGE Publications

De Ramón Fernández FA, Ruiz Fernández D, Prieto Sánchez MT (2019) A decision support system for predicting the treatment of ectopic pregnancies. Int J Med Inform 129:198–204. https://doi.org/10.1016/j.ijmedinf.2019.06.002

Denecke K, Lutz Hochreutener S, Pöpel A, May R (2018) Talking to ana: a mobile self-anamnesis application with conversational user interface. In: International Conference on Digital Health. ACM: New York, US

Devi D, Biswas SK, Purkayastha B (2019) Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique. Conn Sci 31:105–142. https://doi.org/10.1080/09540091.2018.1560394

Diederich S, Brendel AB, Kolbe LM (2019) On conversational agents in information systems research: analyzing the past to guide future work. In: Proceedings of 14th International Conference on Wirtschaftsinformatik. AISel: Siegen, Germany

Dilsizian SE, Siegel EL (2014) Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep 16:441. https://doi.org/10.1007/s11886-013-0441-8

Duan Y, Edwards JS, Dwivedi YK (2019) Artificial intelligence for decision making in the era of big data—evolution, challenges and research agenda. Int J Inf Manag 48:63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

EU (2020) Ethics guidelines for trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation. Accessed 29 Apr 2021

Floridi L, Cowls J, Beltrametti M et al (2018) AI4 people—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach 28:689–707. https://doi.org/10.1007/s11023-018-9482-5

Fox G, James TL (2020) Toward an understanding of the antecedents to health information privacy concern: a mixed methods study. Inf Syst Front. https://doi.org/10.1007/s10796-020-10053-0

Frick N, Brünker F, Ross B, Stieglitz S (2019a) The utilization of artificial intelligence for improving incident management. HMD 56:357–369. https://doi.org/10.1365/s40702-019-00505-w

Frick NRJ, Brünker F, Ross B, Stieglitz S (2019b) Towards Successful Collaboration: Design Guidelines for AI-based Services enriching Information Systems in Organisations. In: Proceedings of the 30th Australasian Conference on Information Systems. ArXiv, Fremantle, Australia, p arXiv:1912.01077

Froomkin AM, Kerr I, Pineau J (2019) Confronting the challenges of the world. Call Holin 61:167–170. https://doi.org/10.2307/j.ctt1p6qpn7.29

Gioia DA, Corley KG, Hamilton AL (2013) Seeking qualitative rigor in inductive research. Organ Res Methods 16:15–31. https://doi.org/10.1177/1094428112452151

Glaser BG (2013) No preconceptions: the grounded theory dictum. Sociology Press, Mill Valley

Gnewuch U, Morana S, Adam M, Maedche A (2017) Towards Designing Cooperative and Social Conversational Agents for Customer Service. In: Proceedings of the Thirty Eighth International Conference on Information Systems. CCBY-NC-ND 4.0 license http://creativecommons.org/licenses/bync-nd/4.0/. South Korea

Gruson D, Helleputte T, Rousseau P, Gruson D (2019) Data science, artificial intelligence, and machine learning: opportunities for laboratory medicine and the value of positive regulation. Clin Biochem 69:1–7. https://doi.org/10.1016/j.clinbiochem.2019.04.013

Gulshan V, Peng L, Coram M et al (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc 316:2402–2410. https://doi.org/10.1001/jama.2016.17216

He J, Baxter SL, Xu J et al (2019) The practical implementation of artificial intelligence technologies in medicine. Nat Med 25:30–36. https://doi.org/10.1038/s41591-018-0307-0

Hebert PC, Meslin EM, Dunn EV (1992) Measuring the ethical sensitivity of medical students: a study at the University of Toronto. J Med Ethics 18:142–147. https://doi.org/10.1136/jme.18.3.142

Hirschauer TJ, Adeli H, Buford JA (2015) Computer-aided diagnosis of Parkinson's disease using enhanced probabilistic neural network. J Med Syst. https://doi.org/10.1007/s10916-015-0353-9

Hulkower R (2010) The history of the hippocratic oath: outdated, inauthentic, and yet still relevant 4 commentary. Einstein J Biol Med 25:41–44

Jain AK, Mao J, Mohiuddin KM (1996) Artificial neural networks: a tutorial. Computer (long Beach Calif) 29:31–44. https://doi.org/10.1109/2.485891

Jiang F, Jiang Y, Zhi H, et al (2017) Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 0:e000101. https://doi.org/10.1136/svn-2017-000101

Kara S, Güven A, Okandan M, Dirgenali F (2006) Utilization of artificial neural networks and autoregressive modeling in diagnosing mitral valve stenosis. Comput Biol Med 36:473–483. https://doi.org/10.1016/j.compbiomed.2005.01.007

Ker J-I, Wang Y, Hajli N (2018) Examining the impact of health information systems on healthcare service improvement: The case of reducing in patient-flow delays in a U.S. hospital. Technol Forecast Soc Change 127:188–198. https://doi.org/10.1016/j.techfore.2017.07.013

Kimani E, Bickmore T, Trinh H, et al (2016) A Smartphone-Based Virtual Agent for Atrial Fibrillation Education and Counseling. In: Lecture Notes in Computer Science: Proceedings of the International Conference on Intelligent Virtual Agents, 10011th edn. Springer: Los Angeles, US. pp 120–127

King A, Bickmore T, Campero M et al (2013) Employing virtual advisors in preventive care for underserved communities: results from the COMPASS study. J Health Commun 18:1449–1464. https://doi.org/10.1080/10810730.2013.798374

Knight W (2017) The dark secret at the heart of AI. MIT Technol Rev. https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/. Accessed 21 Jun

Knijnenburg B, Willemsen M (2016) Inferring capabilities of intelligent agents from their external traits. ACM Trans Interact Intell Syst 6:1–25. https://doi.org/10.1145/2963106

Krittanawong C, Zhang H, Wang Z et al (2017) Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol 69:2657–2664. https://doi.org/10.1016/j.jacc.2017.03.571

Ku CH, Chang Y-C, Wang Y, et al (2019) Artificial Intelligence and Visual Analytics: A Deep-Learning Approach to Analyze Hotel Reviews & Responses. In: Proceedings of the 52nd Hawaii International Conference on System Sciences. HICSS: CC BY-NC-ND 4.0, Hawaii, pp 5268–5277

Larsen KR, Hovorka DS, Dennis AR, West JD (2019) Understanding the elephant: the discourse approach to boundary identification and corpus construction for theory review articles. J Assoc Inf Syst 20:887–927. https://doi.org/10.17705/1jais.00556

Li H, Wang X, Liu C et al (2019a) Dual-input neural network integrating feature extraction and deep learning for coronary artery disease detection using electrocardiogram and phonocardiogram. IEEE Access 7:146457–146469. https://doi.org/10.1109/ACCESS.2019.2943197

Li Y, Deng X, Wang Y (2019b) Introduction to the minitrack on augmenting human intelligence: artificially, socially, and ethically. In: Proceedings of the 52nd Hawaii International Conference on System Sciences, Manoa, Hawaii, pp 5266–5267

Libaque-Sáenz CF, Wong SF, Chang Y, Bravo ER (2020) The effect of fair information practices and data collection methods on privacy-related behaviors: a study of mobile apps. Inf Manag. https://doi.org/10.1016/j.im.2020.103284

López-Martínez F, Núñez-Valdez ER, Lorduy Gomez J, García-Díaz V (2019) A neural network approach to predict early neonatal sepsis. Comput Electr Eng 76:379–388. https://doi.org/10.1016/j.compeleceng.2019.04.015

Luger E, Sellen A (2016) "Like Having a Really Bad PA": the gulf between user expectation and experience of conversational agents. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM: New York, US, pp 5286–5297

Luxton DD (2014a) Recommendations for the ethical use and design of artificial intelligent care providers. Artif Intell Med. https://doi.org/10.1016/j.artmed.2014.06.004

Luxton DD (2014b) Artificial intelligence in psychological practice: current and future applications and implications. Prof Psychol Res Pract 45:332–339. https://doi.org/10.1037/a0034559

Luxton DD (2019) Should watson be consulted for a second opinion? AMA J Ethics 21:131–137. https://doi.org/10.1001/amajethics.2019.131

McKinney SM, Sieniek M, Godbole V et al (2020) International evaluation of an AI system for breast cancer screening. Nature 577:89–94. https://doi.org/10.1038/s41586-019-1799-6

Menai MEB (2015) Random forests for automatic differential diagnosis of erythemato-squamous diseases. Int J Med Eng Inform 7:124–141. https://doi.org/10.1504/IJMEI.2015.068506

Menschner P, Prinz A, Koene P et al (2011) Reaching into patients' homes—participatory designed AAL services: the case of a patient-centered nutrition tracking service. Electron Mark 21:63–76. https://doi.org/10.1007/s12525-011-0050-6

Meuser M, Nagel U (2009) The expert interview and changes in knowledge production. In: AB et al. (eds) Interviewing experts. Palgrave Macmillan, London, pp 17–42

Miles SH (2005) The hippocratic oath and the ethics of medicine. Oxford University Press, New York

Mirbabaie M, Stieglitz S, Brünker F et al (2020) Understanding collaboration with virtual assistants—the role of social identity and the extended self. Bus Inf Syst Eng. https://doi.org/10.1007/s12599-020-00672-x

Mirbabaie M, Stieglitz S, Frick NRJ (2021a) Hybrid intelligence in hospitals: towards a research agenda for collaboration. Electron Mark. https://doi.org/10.1007/s12525-021-00457-4

Mirbabaie M, Stieglitz S, Frick NRJ (2021b) Artificial intelligence in disease diagnostics : a critical review and classification on the current state of research guiding future direction. Health Technol (berl). https://doi.org/10.1007/s12553-021-00555-5

Mitchell T, Cohen W, Hruschka E et al (2018) Never-ending learning. Commun ACM 61:103–115. https://doi.org/10.1145/3191513

Nalini S (2019) Determination of muscles of head acting in whistling. Int J Physiol 7:1. https://doi.org/10.5958/2320-608x.2019.00033.7

Nasirian F, Ahmadian M, Lee O-K (Daniel) (2017) AI-based voice assistant systems: evaluating from the interaction and trust perspectives. In: Proceedings of the Twenty-third American Conference on Information Systems. AISel, Boston, US

Neill DB (2013) Using artificial intelligence to improve hospital inpatient care. IEEE Intell Syst 28:92–95. https://doi.org/10.1109/MIS.2013.51

Page K (2012) The four principles: can they be measured and do they predict ethical decision making? BMC Med Ethics. https://doi.org/10.1186/1472-6939-13-10

Parikh RB, Obermeyer Z, Navathe AS (2019) Regulation of predictive analytics in medicine. Algorithms must meet regulatory standards of clinical benefit. Science 363:6429. https://doi.org/10.1126/science.aaw0029

Pereira C, McNamara A, Sorge L, Arya V (2013) Personalizing public health: your health avatar. J Am Pharm Assoc 53:145–151. https://doi.org/10.1331/JAPhA.2013.12207

Ploug T, Holm S (2020) The right to refuse diagnostics and treatment planning by artificial intelligence. Med Health Care Philos 23:107–114. https://doi.org/10.1007/s11019-019-09912-8

Porra J, Lacity M, Parks MS (2020) "Can computer based human-likeness endanger humanness?"—a philosophical and ethical perspective on digital assistants expressing feelings they can't have. Inf Syst Front 22:533–547. https://doi.org/10.1007/s10796-019-09969-z

Preece A, Webberley W, Braines D et al (2017) Sherlock: experimental evaluation of a conversational agent for mobile information tasks. IEEE Trans Hum Mach Syst 47:1017–1028. https://doi.org/10.1109/THMS.2017.2700625

Price J, Price D, Williams G, Hoffenberg R (1998) Changes in medical student attitudes as they progress through a medical course. J Med Ethics 24:110–117. https://doi.org/10.1136/jme.24.2.110

Qu S, Dumay J (2011) The qualitative research interview. Qual Res Account Manag 8:238–264. https://doi.org/10.1108/11766091111162070

Rai A, Constantinides P, Sarker S (2019) Next-generation digital platforms: toward human-AI hybrids. MIS Q 43:iii–ix

Rauschert S, Raubenheimer K, Melton PE, Huang RC (2020) Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification. Clin Epigenet. https://doi.org/10.1186/s13148-020-00842-4

Reddy S, Fox J, Purohit MP (2019) Artificial intelligence-enabled healthcare delivery. J R Soc Med 112:22–28. https://doi.org/10.1177/0141076818815510

Rezler AG, Lambert P, Obenshain SS et al (1990) Professional decisions and ethical values in medical and law students. Acad Med 65:31–32

Rezler AG, Schwartz RL, Obenshain SS et al (1992) Assessment of ethical decisions and values. Med Educ 26:7–16. https://doi.org/10.1111/j.1365-2923.1992.tb00115.x

Riddick FA (2003) The code of medical ethics of the American Medical Association. https://www.ama-assn.org/delivering-care/ethics/code-medical-ethics-overview. Accessed 21 Jun

Rong G, Mendez A, Bou Assi E et al (2020) Artificial intelligence in healthcare: review and prediction case studies. Engineering 6:291–301. https://doi.org/10.1016/j.eng.2019.08.015

Rosen MA, DiazGranados D, Dietz AS et al (2018) Teamwork in healthcare: key discoveries enabling safer, high-quality care. Am Psychol 73:433–450. https://doi.org/10.1037/amp0000298

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215. https://doi.org/10.1038/s42256-019-0048-x

Salerno J, Knoppers BM, Lee LM et al (2017) Ethics, big data and computing in epidemiology and public health. Ann Epidemiol 27:297–301. https://doi.org/10.1016/j.annepidem.2017.05.002

Schiff D, Borenstein J (2016) AMA Journal of Ethics®. Clin Ethics 18:473–563

Schilling J (2006) On the pragmatics of qualitative assessment. Eur J Psychol Assess 22:28–37. https://doi.org/10.1027/1015-5759.22.1.28

Seeber I, Bittner E, Briggs RO et al (2020) Machines as teammates: a research agenda on AI in team collaboration. Inf Manag 57:103174. https://doi.org/10.1016/j.im.2019.103174

Serrano A, Garcia-Guzman J, Xydopoulos G, Tarhini A (2020) Analysis of barriers to the deployment of health information systems: a stakeholder perspective. Inf Syst Front 22:455–474. https://doi.org/10.1007/s10796-018-9869-0

Shaked N (2017) Avatars and virtual agents—relationship interfaces for the elderly. Healthc Technol Lett 4:83–87. https://doi.org/10.1049/htl.2017.0009

Siau K, Wang W (2018) Building trust in artificial intelligence, machine learning, and robotics. Cut Bus Technol J 31:47–53

Siau K, Wang W (2020) Artificial intelligence (AI) ethics. J Database Manag 31:74–87. https://doi.org/10.4018/jdm.2020040105

Sonja M, Ioana G, Miaoqing Y, Anna K (2018) Understanding value in health data ecosystems: a review of current evidence and ways forward. Rand Health Q 7(2):3, PMID: 29416943; PMCID: PMC5798965

Ting DSW, Cheung CYL, Lim G et al (2017) Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. J Am Med Assoc 318:2211–2223. https://doi.org/10.1001/jama.2017.18152

Vayena E, Blasimme A, Cohen IG (2018) Machine learning in medicine: addressing ethical challenges. PLoS Med 15:4–7. https://doi.org/10.1371/journal.pmed.1002689

vom Brocke J, Simons A, Niehaves et al (2009) Reconstructing the giant: on the importance of rigour in documenting the literature search process. In: Proceedings of the 17th European Conference on Information Systems, AISel, Verona, Italy

vom Brocke J, Simons A, Riemer K et al (2015) Standing on the shoulders of giants: challenges and recommendations of literature search in information systems research. Commun Assoc Inf Syst. https://doi.org/10.17705/1CAIS.03709

Walker S (2020) Poland delays abortion ban as nationwide protests continue. https://www.theguardian.com/world/2020/nov/03/poland-stalls-abortion-ban-amid-nationwide-protests. Accessed 21 Jun

Wang Y (2020) Toward an understanding of responsible artificial intelligence practices. In: Proceedings of the 53rd Hawaii International Conference on System Sciences, HICSS CC BY-NC-ND 4.0, Hawaii, pp 4962–4971

Wears RL, Berg M (2005) Computer technology and clinical work. J Am Med Assoc 293:1261–1263. https://doi.org/10.1001/jama.293.10.1261

Yu KH, Kohane IS (2019) Framing the challenges of artificial intelligence in medicine. BMJ Qual Saf 28:238–241. https://doi.org/10.1136/bmjqs-2018-008551

Yu KH, Beam AL, Kohane IS (2018) Artificial intelligence in healthcare. Nat Biomed Eng 2:719–731. https://doi.org/10.1038/s41551-018-0305-z

Zeng Y, Lu E, Huangfu C (2019) Linking Artificial Intelligence Principles. In: AAAI Workshop on Artificial Intelligence Safety. arXiv, Honolulu, Hawaii. https://arxiv.org/abs/1812.04814. Accessed 21 Jun

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paper 5: Understanding Collaboration with Virtual Assistants – The Role of Social Identity and the Extended Self

Type (Ranking, Impact Factor)	Journal article (B, 7.9)
Status	Published
Rights and permissions	Open access
Authors	Mirbabaie, M., Stieglitz, S., Brünker, F., Hofeditz, L., Ross, B., & Frick, N. R. J.
Year	2021
Outlet	Business & Information Systems Engineering (BISE)
Permalink / DOI	https://doi.org/10.1007/s12599-020-00672-x
Full citation	Mirbabaie, M., Stieglitz, S., Brünker, F., Hofeditz, L., Ross, B., & Frick, N. R. J. (2021). Understanding Collaboration with Virtual Assistants – The Role of Social Identity and the Extended Self. *Business and Information Systems Engineering*, 63(1), 21–37. https://doi.org/10.1007/s12599-020-00672-x.

Bus Inf Syst Eng 63(1):21–37 (2021)
https://doi.org/10.1007/s12599-020-00672-x

RESEARCH PAPER

Understanding Collaboration with Virtual Assistants – The Role of Social Identity and the Extended Self

Milad Mirbabaie · Stefan Stieglitz · Felix Brünker · Lennart Hofeditz · Björn Ross · Nicholas R. J. Frick

Received: 15 May 2020 / Accepted: 20 October 2020 / Published online: 1 December 2020

Abstract Organizations introduce virtual assistants (VAs) to support employees with work-related tasks. VAs can increase the success of teamwork and thus become an integral part of the daily work life. However, the effect of VAs on virtual teams remains unclear. While social identity theory describes the identification of employees with team members and the continued existence of a group identity, the concept of the extended self refers to the incorporation of possessions into one's sense of self. This raises the question of which approach applies to VAs as teammates. The article extends the IS literature by examining the impact of VAs on individuals and teams and updates the knowledge on social identity and the extended self by deploying VAs in a collaborative setting. Using a laboratory experiment with N = 50, two groups were compared in solving a task, where one group was assisted by a VA, while the other was supported by a person. Results highlight that employees who identify VAs as part of their extended self are more likely to identify with team members and vice versa. The two aspects are thus combined into the proposed construct of virtually extended identification explaining the relationships of collaboration with VAs. This study contributes to the understanding on the influence of the extended self and social identity on collaboration with VAs. Practitioners are able to assess how VAs improve collaboration and teamwork in mixed teams in organizations.

Keywords Virtual collaboration · Virtual assistants · Social identity theory · Extended self · Information systems · Organizations · Virtually extended identification

1 Introduction

In virtual collaboration, teams are required to collaborate via technology (de Vreede and Briggs 2005; Changizi and Lanz 2019) which can result in a lack of a common social identity (Vahtera et al. 2017). With some technologies, such as virtual assistants (VAs), the role of technology is changing from a mere tool for virtual collaboration with other humans to its own virtual collaboration with VAs (Maedche et al. 2019; Seeber et al. 2020a). VAs are software programs that can be addressed via voice or text commands and respond to the users' input (Brachten et al. 2020). They are increasingly being used in organizations to optimize internal processes by assisting in the execution of work-related tasks (Norman 2017) to achieve, for example, increased customer satisfaction, thus creating substantial advantages over competitors (Benbya and Leidner 2018; Yan et al. 2018). Unlike physical robots, such as Nao or Pepper, which have a physical human representation

Accepted after two revisions by the editors of the special issue.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12599-020-00672-x) contains supplementary material, which is available to authorized users.

M. Mirbabaie (✉)
Faculty of Business Studies and Economics, University of Bremen, Enrique-Schmidt-Straße 1, 28359 Bremen, Germany
e-mail: milad.mirbabaie@uni-bremen.de

S. Stieglitz · F. Brünker · L. Hofeditz · N. R. J. Frick
Faculty of Engineering, Department of Computer Science and Applied Cognitive Science, Professional Communication in Electronic Media/Social Media, University of Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany

B. Ross
School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom

(Maniscalco et al. 2020), a physical interaction with VAs is not possible. However, VAs are used in virtual collaboration (Seeber et al. 2020a; Panganiban et al. 2020). It is predicted that they will be used by at least a quarter of employees working in virtual teams within the next two years (Maedche et al. 2019). To understand virtual collaboration between humans and machines such as VAs, knowledge from human-to-human collaboration research should be exploited (Demir et al. 2020).

Nowadays, many team members, such as those in global virtual project teams (Massey et al. 2003), are physically widely distributed and collaborate primarily virtually (Plotnick et al. 2016; Hassell and Cotton 2017; Andres and Shipps 2019). Virtual collaboration ranges from working together in virtual computer-generated worlds (Franceschi et al. 2009; Kohler et al. 2011) to collaboration using tools such as Google Drive (Van Ostrand et al. 2016). Successful virtual collaboration is influenced by aspects such as social presence (Franceschi et al. 2009) and social identity (Lin 2015; Vahtera et al. 2017). Identifying with team members at the workplace as a social group contributes significantly to improving the individual performance of each employee and encourages achieving an overarching goal more efficiently (Lin 2015; Porck et al. 2019). One's own identity can partially be depicted within the framework of a virtual collaboration, for example, by visualizing gender, age, and social class via embodiment through an avatar (Schultze 2010). The social identity of team members can also be transferred to virtual collaboration (Guegan et al. 2017). Social identity describes the identification with other (virtual) team members and the maintenance of one's own identity by comparing one's self-concept with other people's perceived values, norms, and characteristics (Brown 2000).

Research on the role of VAs as team members is not a recent development (Seeber et al. 2020a; Panganiban et al. 2020; Demir et al. 2020). However, it is still largely unexplored whether VAs are perceived as part of one's team or as a simple tool or object in virtual collaboration. The identification with an object as part of one's self has been called the "extended self" (Belk 1988; Tian and Belk 2005; Clayton et al. 2015) and has been transferred to the workplace and the digital world. People extend their identity by incorporating capabilities that fit to their self-concept, and thus, positively enhance their self.

In contrast, the theory of social identity focuses on the comparison with other humans in order to form and maintain one's identity (Tajfel and Turner 1986). This apparent contradiction raises the question of which approach applies to VAs as team members in virtual collaboration. Examining this is fundamental to understand how and with what purpose VAs should be deployed in organizations as collaborative partners. Deploying VAs could help organizations to save valuable resources when they are used as tools to assist employees in work-related tasks or when they behave as team partner in order to increase team identity and therefore team efficiency. To examine the role of VAs in virtual collaboration in detail, our research is guided by the following research question:

> How does identification with VAs vs. that with humans as virtual team members differ in virtual collaboration?

To answer the research question, we conducted a laboratory experiment with 50 participants. Those in the experimental group were asked to solve a typical work-related task in collaboration with a text-based VA, while the control group was assisted by another human via chat. We measured and compared the extended self and the social identity for both groups as well as the perceived workload. This paper contributes to research and practice by extending our understanding of the collaboration between employees and VAs in an organizational context to drive future research in this field of high relevance. Information systems (IS) researchers will find the insights helpful to understand what influence the extended self and social identity theory have on virtual collaboration with VAs assisting in work-related tasks. To guide future research, we introduce the concept of virtually extended identification as a combination of social identity and the extended self for virtual collaboration between VAs and employees.

2 Related Work: Virtual Assistants in Organizations

Collaboration technologies have a long history in IS research (Schwabe 2003; Frohberg and Schwabe 2006; Bajwa et al. 2007; You and Robert 2018). For VAs, as one of these technologies, the IS community uses a variety of definitions (e.g., Maedche et al. 2019; Seeber et al. 2020a; Diederich et al. 2020). Luger and Sellen (2016) define CAs as "*IS that enable the interaction with users via natural language.*" Stieglitz et al. (2018) state that VAs in enterprises "*can be addressed* via *voice or text and that can respond to the users input (i.e. assist) with sought-after information.*" VAs can generally be explained as software programs that can be addressed via different modes of communication (e.g., written or spoken natural language), assisting with tasks or executing them autonomously (Brachten et al. 2020). Related terms include but are not limited to chatbots (Stieglitz et al. 2018), conversational agents (Diederich et al. 2020), and digital assistants (Maedche et al. 2019). Research divides the concept of VAs into various categories, such as design characteristics or assistance domain (Knote et al. 2019). However, systems are usually classified along two dimensions (Gnewuch et al. 2017) – their primary mode of communication (e.g., text-based or speech-based) (Lee et al. 2009) and their main purpose (narrow or broad task) (Nunamaker et al. 2011). A categorization into one of these classes is not always possible due to potential overlaps. For example, VAs can be augmented to cope with individual requirements (Chung et al. 2017), and text-based systems might convert human language into text to process information (Gnewuch et al. 2017).

VAs need to be differentiated from a number of related concepts. VAs can distinguish among and interpret the emotions of individuals within teams (McDuff and Czerwinski 2018) and use different language styles to adapt to varying users (Gnewuch et al. 2020). Thereby they might use social cues, including the dimensions of verbal (e.g., jokes, temporal expressions, or self-disclosure), visual (e.g., emoticons, facial expressions, or agent visualization), auditory (e.g., voice gender, grunt, and moan or laughing), and invisible (e.g., first turn, response time, or tactile touch; Feine et al. 2019). Thus, collaborating with VAs might not be restricted to certain commands, phrases, or keywords; rather, individuals can use their habitual language (McTear 2017; Feine et al. 2019). Although VAs theoretically have various verbal, visual, auditory and invisible characteristics that can impact social behavior in humans (Feine et al. 2019), in practice it is still hardly possible to simulate fully human behavior. VAs are usually capable of supporting a narrow task (Davenport 2018), but may not be able to provide appropriate answers in every context. They are therefore usually characterized by a certain selection of social cues, but cannot represent a fully human consciousness (Russel and Norvig 2016).

The ongoing improvements to artificial intelligence (AI) and machine learning (ML) algorithms as a prerequisite to developing collaborative systems had led to an increasing concentration on VAs as work facilitators (Berg et al. 2015; Spohrer and Banavar 2015; Luger and Sellen 2016; Knijnenburg and Willemsen 2016; Nasirian and Ahmadian 2017). The use of VAs in organizations is valuable for facilitating internal processes and supporting employees in better completing their tasks as well as generating additional revenue or cost savings (Quarteroni 2018). VAs are used for direct interaction with consumers, and they positively affect customer satisfaction (Verhagen et al. 2014). Question-and-answer assistants facilitate onboarding processes of new hires (Shamekhi et al. 2018). The workload of employees is reduced by supporting the resolution of customer incidents (McTear 2017) and the execution of work-related tasks (Brachten et al. 2020).

Current research demonstrates that VAs can improve virtual collaboration (Waizenegger et al. 2020; Seeber et al. 2020a). Organizational human teams frequently fall short of their possibilities (Kozlowski and Ilgen 2007), thus the use of a VA as a legitimate virtual team member and socio-technical ensemble (Seeber et al. 2018) might foster decision making and improve team collaboration (Waizenegger et al. 2020; Seeber et al. 2020b). The integration of VAs as virtual colleagues is valuable to increase the effectiveness of virtual collaboration in teams (Goodbody 2005). With their unique characteristics (Maedche et al. 2019; Feine et al. 2019) and ongoing application in practice (Brachten et al. 2020), it can be assumed that an increasing degree of team dynamics from purely human virtual teams can be transferred to human–machine teams.

3 Theoretical Background

3.1 Social Identity

Social identity is a grounded concept that can influence the performance of virtual teams (Lin 2015). In social identity theory, Tajfel and Turner (1986) assume that human identity is not only composed of individually unique character traits and physical characteristics but also of belonging to certain social groups. This might include people of the same age group, family, friends, and even work colleagues (Bartels et al. 2019).

By comparing with other social groups, such as other departments or competing organizations, individuals try to draw a line to better understand who they themselves are (Tajfel and Turner 1986). People, such as employees, try to differentiate from others by means of positive characteristics that they attribute to themselves, which is known as intrinsically motivated positive distinctiveness (Haslam 2004). At the workplace, such characteristics can be team cohesion or quality of work.

In IS research, social identity theory at the workplace has been considered from perspectives including the psychological (Pepple and Davies 2019; Klimchak et al. 2019), the organizational (Dahling and Gutworth 2017; Mueller et al. 2019), and the societal viewpoints (Kenny and Briner 2013).

However, most previous studies have focused on examining social identity in human-to-human collaboration and the resulting social behavior (Kohler et al. 2011). With technologies such as VAs, which are capable of utilizing human social cues (Maedche et al. 2019), the role of technology is changing, and the boundaries between people and technology are blurring (Pickard et al. 2013). According to Young-Jae et al. (2020), people perceive it as increasingly difficult to describe the uniqueness of humans compared to machines and AI as the technology itself could be perceived as a social actor (Wang 2017; Edwards et al. 2019). This actor is less a technological environment than a possible new individual that could be part of an in-group or out-group in the context of social identity formation.

Revealing insights about the relationship between people and AI will open up new opportunities for organizations and interesting insights for further research. However, social identity theory is not the only concept that could explain the role of AI in virtual collaboration. Another concept from psychology addressing the social relationship between humans and objects (e.g., technologies) could also help to better understand the virtual collaboration between humans and machines – the extended self (Belk 2013).

3.2 The Extended Self

People develop and maintain several identities according to the context of their current situation (Burke 2006). Thus, Burke and Stets (2009) argue that people play different roles. For example, people face specific actors and topics at the workplace according to the situation, such as a team meeting or an idea pitch. Likewise, people need to adapt to other situations at home, such as in the context of the education of one's children. Individuals have various roles prepared for the unique situations they face. Besides those roles, people maintain only one underlying self-concept connected to fundamental rules and values that they develop over time by categorizing in relation to others (Stets and Burke 2000; Burke and Stets 2009). Hence, identity is a well-discussed research area connected to various disciplines, such as psychology (Tajfel and Turner 1986), social psychology (Leary and Tangney 2011), sociology (Stets and Biga 2003), and economic psychology (Belk 1988). However, it is worth analyzing identity in relation to the increasing role of information technology as a new resource in our life and work (Tian and Belk 2005; Carter et al. 2015).

People extend their selves by considering particular possessions in order to supplement their self (Belk 1988, 2013). However, the concept of possessions is not limited to external objectives; it can also include other people or group possessions. Furthermore, under the perspective of upcoming technology, Belk (2013) argues that people can also consider digital possessions as potential extensions of the self. This might be achieved by, for example, dematerialization, sharing, or distributed memories. Particularly in the workplace of technology organizations, Tian and Belk (2005) argue that employees need to decide which part of the self fits the current situation of the work, and how. On one hand, this decision includes the process of negotiations between the "me" and the situation. On the other hand, this decision may stay hidden or might be retracted.

However, due to the integral role of information technology in everyday life and work, understanding information technology, for example, in the form of virtual collaboration and new social actors such as VAs, has become a relevant endeavor for IS research (Carter et al. 2015). In this regard, maintaining and extending the self are two central functions in the context of information technology and identity (Carter and Grover 2015). It is necessary to answer the question "Who am I in relation to this technology?" (Vignoles et al. 2011; Carter et al. 2015). This material perspective focuses on individual thinking and behavior (Dittmar 2011). Therefore, material identities are verified when people gain control and mastery of an object that they are interacting with.

Furthermore, people have a fundamental need to expand the self and seek self-enhancement. They can achieve this by supplementing social or physical resources, perspectives, and identities (Aron et al. 2003). One possible way for people to achieve this enhancement is by consolidating capacities yielded by (material) objects to which they have become emotionally attached (Belk 1988, 2013; Carter et al. 2015).

3.3 Derivation of Hypotheses

Social identity theory and the extended self describe two alternative pathways to maintain and form an individual's identity (Tajfel and Turner 1986; Belk 1988, 2013; Stets and Burke 2000). Social identity theory holds that identification with other (social) actors leads to a sense of belonging to the group (external attribution of an actor's values to the self; Tajfel and Turner 1986; Stets and Burke 2000). In comparison, the perspective of the extended self conceptualizes that a positive identification with an (virtual) object leads to an association of capabilities, characteristics, or meanings directly to the self (internal attribution of an actor's values to the self; Belk 1988, 2013; Tian and Belk 2005). Based on the considerations of the theoretical background, Table 1 contrasts how the extended self and social identity determine the perception of a VA as a team member.

Previous research has stated that VAs can change how we live and how we work (Wang and Siau 2018; Dias et al. 2019); thus, employees and organizations need to find out how to collaborate with VAs within their virtual teams (Seeber et al. 2018). People spend a large part of their lives at their workplaces, where they build and maintain complex social relationships (Ellemers 2004). Their work and team colleagues hence represent important social resources through which individuals build their social identity and develop in-group and out-group behaviors (Tajfel and Turner 1986). Thus, questions arise as to whether VAs are perceived as part of these social resources, and whether

Table 1 Social identity theory and the extended self in virtual collaboration with VAs

Perception of VAs as virtual team members	Confirmation of self-concept	Contradiction of self-concept
Social identity theory	The VA is perceived as a social actor. Perceived values, rules, and standards also apply to the self. This leads to a sense of belonging to the group/person (Tajfel and Turner 1986; Stets and Burke 2000; Edwards et al. 2019)	The VA is perceived as a social actor. Perceived values, rules, and standards disaccord with the self. This leads to a dissociation from the group/person (Tajfel and Turner 1986; Stets and Burke 2000; Edwards et al. 2019)
Extended self	The VA is perceived as part of the self. Capabilities, attributes, or associations of the VA are attributed to the self (Belk 1988, 2013; Burke 2006; Carter and Grover 2015)	The VA is not perceived as part of the self to protect the self-concept. Capabilities, attributes, or associations of the VA are not attributed to the self (Belk 1988, 2013; Burke 2006; Carter and Grover 2015)
Similarities	Considering perceived aspect, such as values, rules, capabilities, and attributes of the VA that fit positively with the individual's self	Dissociation of perceived aspect, such as values, rules, capabilities, and attributes of the VA that do not fit with the individual's self

they influence the identity of employees remains unanswered. As most VAs are designed as supportive tools (Lamontagne et al. 2014) and not as equivalent virtual team members, they still remain IS (Luger and Sellen 2016). Therefore, it can be assumed that collaborating with a VA as a chat partner or with a human chat partner impacts the identification with that chat partner. We therefore developed the following hypothesis:

H1: Virtually collaborating with a VA or a human chat partner impacts the identification with the chat partner.

VAs can increase collaboration within virtual teams (Bittner et al. 2019; Seeber et al. 2020a). However, when employees use VAs as supportive tools for solving work-related tasks, it is likely that they interact less with their virtual human team partners. Nevertheless, the time employees spend with their virtual team impacts the team identification (Massey et al. 2003). Therefore, we derived the following hypothesis:

H2: Identification with the human team is lower after collaborating with a VA than before.

Furthermore, Carter et al. (2012) have shown that young students extended their self-concepts by including the capabilities of their smartphones. According to Tian and Belk (2005) as well as Belk (2013), also digital tools or technology might be considered as part of one's extended self. This identification and enhancement might also be attained by using, and thus incorporating, the capabilities of a VA in a certain context, such as virtual collaboration at the workplace. It remains unclear whether a new technology such as a VA will be perceived as part of one's extended self. Thus, we derived the following hypothesis:

H3: Virtually collaborating with a VA or a human chat partner impacts the perception of the respective collaboration partner as part of one's extended self.

Research has shown that VAs are perceived as supportive technology (Brachten et al. 2020). However, it still needs to be researched what role such technology plays in self-identification at the workplace. Regarding social identity theory and extended self, two alternative pathways appear to maintain and form an individuals' identity (Tajfel and Turner 1986; Belk 1988). According to social identity theory, identification with other (social) actors leads to a sense of belonging to the group. Those social actors could be human team members or VAs (Edwards et al. 2019). However, perceiving VAs as social actors (Edwards et al. 2019) may contradict the perception of VAs as technology (Lamontagne et al. 2014; Carter et al. 2015). Therefore, it is possible that the approaches of social identity and the extended self interfere in virtual collaboration with VAs. Based on these assumptions, we derive that individuals' identification with the team contradicts their identification with technology as a part of their extended self. We, therefore, derive the following hypothesis:

H4: The individual's identification with the team negatively correlates with the individual's identification with technology as a part of their extended self.

4 Method

4.1 Participants

In this study, we conducted a laboratory experiment to examine how VAs in virtual teams are perceived when they assist individuals in performing tasks. The experiment was conducted in a lab at a German university between

November 12, 2019 and February 10, 2020. We invited people via email, social network sites, and direct contact. Participation was voluntary and could be terminated without providing any reasons. As prerequisites, participants had to be at least 18 years old and experienced in teamwork within an organization. In total, 50 people took part in our study. We randomly assigned the participants into two groups, resulting in a well-balanced sample of 25 participants for each condition. The groups were formed ensuring that the proportion of women and men was approximately equal by frequently checking the distribution of gender across groups. If the distribution of subjects was skewed, the smaller group was prioritized. However, due to extreme responding indicating a response bias, we excluded four participants from the total sample. This yielded a total of 46 participants (24 in the VA group). In the control group, the participants were asked to perform a task with the help of a human chat partner. In our experimental group, the participants were asked to solve the same task using a VA. In both cases, the collaboration with the counterpart was possible via the online chat platform Slack.[1] In both groups, a trained experimenter supervised the subjects to secure the subjects' attention during the course of the study. Overall, 84% of the participants were female (N = 39), and ages ranged from 18 to 63 (M = 23.1, SD = 7.54). Furthermore, 73% of the participants had passed the equivalent of their A-levels, while 15% held a bachelor's degree.

4.2 Materials

For our lab experiment, we used a set of questionnaires and modified scales to measure the constructs of interest. These were composed of questions on the extended self, social identity theory, demographic data, perceived workload, satisfaction, and the evaluation and perception of the VA. The analyses were calculated using the software tools Jamovi (1.0.8.0) and SPSS Statistics (Version 25). All data were presented and gathered via the LimeSurvey interface (Version 3.17.5).

4.2.1 Virtual Assistant

To examine how social identity is influenced and whether a VA expands one's own self, we developed a text-based system with the help of Google's cloud service DialogFlow.[2] By using underlying ML technologies, this platform provides easy access to the development of natural and rich conversational interfaces (Canonico and Russis 2018).

To keep the interaction with the VA as simple as possible, we developed a system using a text-based interface (Araujo 2018), which was integrated into the online chat platform Slack, one of the most widespread systems for simplified organizational communication. Participants were able to interact with the VA simply by using a keyboard and computer screen (cf. Fig. 1). We explicitly avoided using further influential factors, such as voice commands or embodied avatars, to keep the interaction straightforward. Moreover, embodiment does not necessarily affect social behavior (Schuetzler et al. 2018). The VA supported the participants in handling the task by providing answers based on distinct keywords to questions posed. The feedback included a question–answer component (Morrissey and Kirakowski 2013; Lamontagne et al. 2014), which could be queried to gain information, support, and instruction about the specific task. However, the VA is only able to support the user in solving the ask by giving applicable hints but does not provide an actual solution for the task.

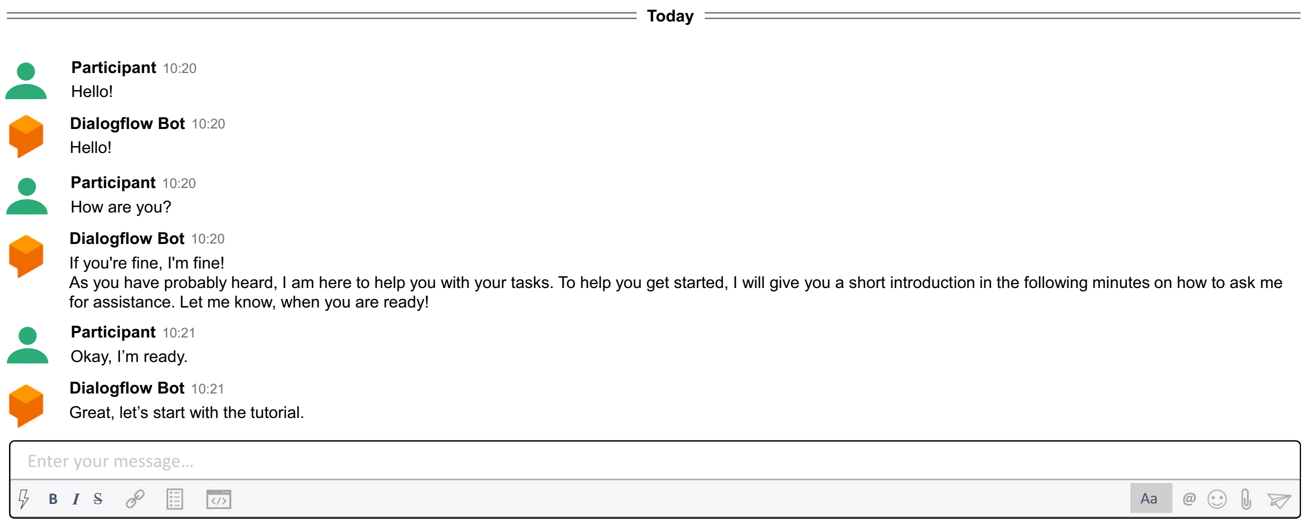
We deliberately chose aspects such as response time to be comparable between both groups to reduce potential influences on the performance and identification with the team member (Massey et al. 2003). Furthermore, the name of the VA (DialogFlow Bot) directly points to a VA as a collaboration partner. Therefore, the subjects should be aware that they were interacting with either a human or a VA. Although our VA had basic conversational skills and social cues such as 'Ask to start', 'Tips and advice', 'Excuse' or 'Greeting and farewell' (Feine et al. 2019) we did not aim to differ specific social cues between the VA and the human (Feine et al. 2019), because that was not our research focus.

We aimed to provide a medium level of social cues to ensure that the VA does not influence the results in one specific direction. Implementing more social cues may favor the perception of the VA as a social actor. In contrast, less social cues could increase the probability of perceiving the VA as a technical tool. With this, we ensured that potential differences in the perception of the team member are due to the team member's nature (VA or human). To summarize, the goal is not to deceive the subjects about the chat partner but to investigate the difference in perception of the VAs and humans based on the subject's awareness about the chat partner.

To ensure that the given task is realistic but manageable during the experiment, we conducted a pre-study to verify its suitability. This approach also served as verification of the operability of the VA to guarantee a seamless collaboration during the experiment. The test was performed with a sample of 10 students (6 female, 4 male) with ages ranging from 22 to 31 (M = 25), which were randomly selected at a university. We compared a text-based task

[1] https://slack.com/.

[2] https://dialogflow.com/.



Participant 10:20
Hello!

Dialogflow Bot 10:20
Hello!

Participant 10:20
How are you?

Dialogflow Bot 10:20
If you're fine, I'm fine!
As you have probably heard, I am here to help you with your tasks. To help you get started, I will give you a short introduction in the following minutes on how to ask me for assistance. Let me know, when you are ready!

Participant 10:21
Okay, I'm ready.

Dialogflow Bot 10:21
Great, let's start with the tutorial.

Fig. 1 Example of a chat interaction between a participant and the VA

(TBT) with the critical path method (CPM). The TBT required participants to read texts about topics that do not rely on previous knowledge. In contrast, the CPM sorts activities according to their dependencies and logical order for determining the overall duration. Both tasks are commonly performed in organizations. The time limit for the execution was 10 minutes. We measured the perceived workload using the NASA Task Load Index (NASA-TLX). On average, participants given the CPM task achieved higher NASA-TLX scores (M = 12.5, SD = 3.85) than the TBT group (M = 6.36, SD = 4.06). This difference of 6.13 was significant (95% CI [0.35, 11.91], t (8) = 2.44, $p = 0.040$). Furthermore, it represents a large effect, $d = 0.98$. We assess the CPM task to be more demanding of participants compared to the TBT. Hence, participants benefit more from a VA when being assisted with the CPM, justifying its choice for the experiment.

4.2.2 *Social Identity*

We used two different questionnaires to measure collective social identity as well as personal identification with the team. For identification with the team, we used the About-Me Questionnaire (Maras et al. 2018), in which the respondents were first asked to indicate how much they felt they belonged to the social group at their workplace. This questionnaire consists of four items, which are rated on a five-point Likert scale. One example item was "*I like being with my team.*" The subscale of the About-Me Questionnaire had a medium-to-high reliability for the first ($\alpha = 0.759$) and second ($\alpha = 0.732$) measurement time points. The About-Me Questionnaire was queried both before and after the interaction with the chat partner to determine a possible change of the specific social identity.

In addition to the two measurement time points, we asked whether in the interaction the VA or human chat partner was perceived as part of the social group at work. This took place after the chat interaction. For this purpose, we used a modified About-Me Questionnaire (Identification with the chat partner). An example item was "*I am similar to my virtual assistant.*" We decided to use the scale directed toward the chat partner to check for possible differences between the general social identity attitude and the social identity attitude toward the interaction scales. The subscale of the modified About-Me Questionnaire had a high reliability, $\alpha = 0.835$.

4.2.3 *The Extended Self*

To measure the extended self, we used the extended self scale by Sivadas and Machleit (1994). The scale is largely based on Belk's (1988) view of the extended self. With the scale, Sivadas and Machleit (1994) aimed to assess the degree of incorporation of possessions into the extended self. The scale consists of six components scored on a seven-point Likert scale. The subscale of the general extended self scale (GES) had high reliability, $\alpha = 0.839$. We chose the scale as it was feasible to adopt for a VA as the considered object for the items. After the chat interaction with the VA or the human, the participants had to answer an adapted version of the extended self scale (AGES) related to the specific chat partner. The AGES measures to what extent the subjects perceiving the chat partner as part of one's self. An example item was "*My virtual assistant is part of what I am.*" The subscales of the second measurement scored a high reliability, $\alpha = 0.886$.

4.2.4 NASA-TLX

To determine the perceived workload of the task, we used the NASA-TLX (Galy et al. 2012), a valid measurement developed by the National Aeronautics and Space Administration (NASA; Hart and Staveland 1988). Examining the perceived workload is important to check whether the new VA influences the performance due to the potential need for increased cognitive resources to interact with a new technology. This assessment tool has successfully been used in several research approaches and proven to be valuable for laboratory experiments (Rubio et al. 2004; Noyes and Bruneau 2007; Cao et al. 2009). The NASA-TLX includes the following six subjective subscales: (1) mental demand, (2) physical demand, (3) temporal demand, (4) performance, (5) effort, and (6) frustration (Hart 2006, p. 904). Mental demand explains how much cognitive activity is needed, and physical demand, in contrast, explains how much manual activity is needed. Temporal demand represents the perceived time pressure. Performance describes the perception of one's own personal accomplishment, effort is the opinion of how much work had to be done to reach a result, and frustration refers to the level of disappointment during the execution of a task. The subscale scored a high reliability, α = 0.808.

4.2.5 Satisfaction

To analyze the perceived satisfaction of the chat interaction via the communication interface, we used the possession satisfaction index (PSI) by Scott and Lundstrom (1990). Measuring the perceived satisfaction may allow us to reveal potential influences that could be caused by the individual perception of the interaction. The PSI uses a seven-point semantic differential scale and contains of three two-pole items of (1) satisfied/dissatisfied, (2) pleased/displeased, and (3) favorable/unfavorable. Furthermore, the PSI scored a high reliability, α = 0.924.

4.3 Procedure

We divided our experiment into one experimental group and one control group. Both groups were alternately tested and told that they should consider the situation as if they were at a workplace they are used to. In the experimental condition, we requested the participants to solve a task in collaboration with a VA. In the control condition, we replaced the collaboration partner with a human chat partner. The procedure of the experiment followed the structure described in the following. All major steps of our experiment are visualized in Fig. 2.

Fig. 2 Main steps of the conducted procedure

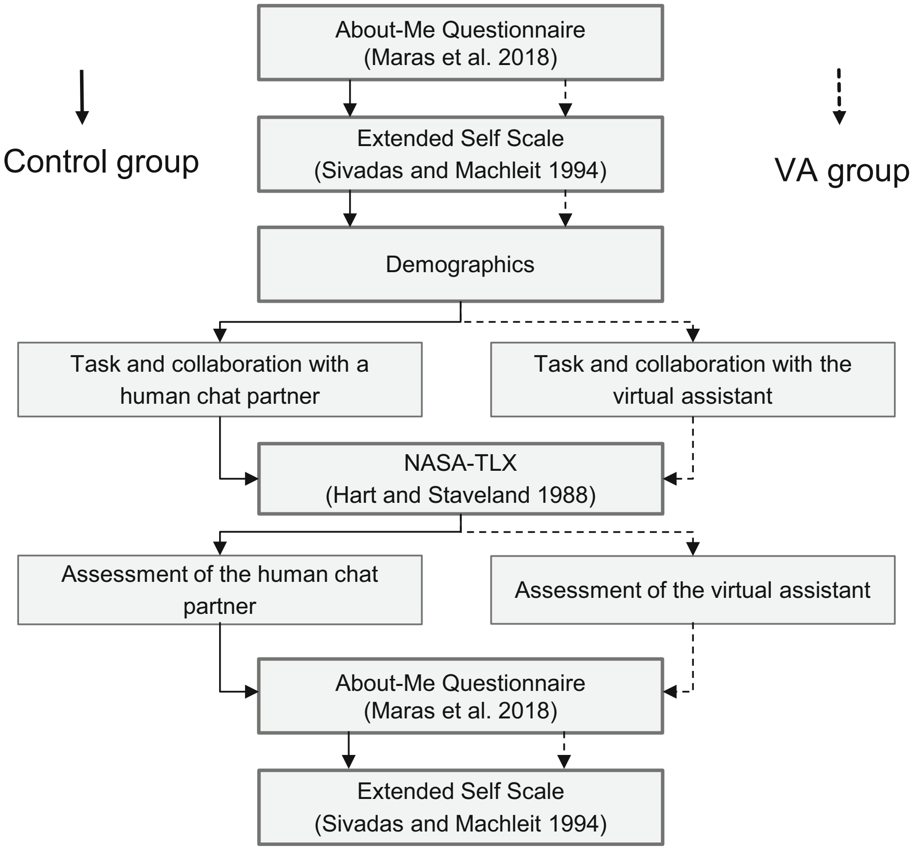

About-Me Questionnaire (Maras et al. 2018)
Control group
VA group
Extended Self Scale (Sivadas and Machleit 1994)
Demographics
Task and collaboration with a human chat partner
Task and collaboration with the virtual assistant
NASA-TLX (Hart and Staveland 1988)
Assessment of the human chat partner
Assessment of the virtual assistant
About-Me Questionnaire (Maras et al. 2018)
Extended Self Scale (Sivadas and Machleit 1994)

Table 2 Correlations in the VA group between perceived satisfaction and the single NASA-TLX items

Items NASA-TLX	N	r	*p*
Performance	24	.575	.003
Effort	24	.506	.012
Frustration	24	534	.007

Comparing the achieved score in the CPM task between the human group (M = 15.2, SD = 6.13) and the VA group (M = 17.2, SD = 6.84) revealed no significant difference, *p* = .359 and *d* = .315

First, we briefed the participants about the experiment. Furthermore, we asked them to read an introductory text and to start with the survey. We reminded the participants that they should imagine they are in a normal working situation and that they should relate the questions to the perception of their current team at work. Initially, we had administered general questionnaires on the extended self, social identity theory, and demographic data. In addition to demographic data such as age, gender, and educational level, we also collected information about the current professional activity and the industry in which the respondents are currently working.

After that, we asked both groups to solve a CPM. To compare performance between the groups, we awarded a point for each correct path and node. This yielded a maximum achievable score of 28. The goal was to plan a research project for a market research unit of a large company. Participants had to arrange an unordered list with various process steps (such as "develop study idea," "literature research," "conducting the study," and "develop methodology") to identify the minimal throughput time. They were to read an introductory text and an example to gain a rough understanding of the task, and we told them that they would have to solve a similar task shortly.

We informed the experimental group that they would have the support of a VA who is well versed with the CPM, whereas we told the control group that they would be contacting a human chat partner. The VA as well as the human chat partner could be contacted via a Slack chatroom. To familiarize them with the interaction, we instructed the participants to introduce themselves to the assistant (or human chat partner), whereby the assistant (or human chat partner) guided them through a tutorial dialog. After this familiarization phase, we provided the CPM task, which the participants had to solve within ten minutes. We advised them to contact the VA (or human chat partner) when any questions arose. We designed the task in such a way that the participants did not have all the necessary information for the required solution in advance in order to initiate interactions with the VA. After ten minutes of processing time, the examiner received the solution. We then requested that the participants continue the survey. With the following questions, we aimed to evaluate the assistant and assess their skills during the task. Subsequently, we enquired the questionnaires on social identity theory and extended self a second time to determine a possible difference in perception. After completion of the last question, we provided a short written debriefing to the respondents to explain what had been examined in the study.

To counteract possible disruptive factors that can arise from interaction with a real human in the control group, the human chat partners followed a semi-structured guideline to ensure that the information provided was as similar as possible to that of the VA. The chat partners were controlled by one experimenter, who switched to the adjoining room for both conditions.

4.4 Influence of the Perceived Workload, Satisfaction, and Demographics on the Groups

To ensure that the results would not be unduly influenced by further variables such as the age, gender, or education of the participants or satisfaction with the chat interaction or the perceived workload, we conducted the following analyses. Determining demographical influences on the main constructs of the study revealed no significant correlation between age and gender and the extended self and social identity scales. However, we observed a small correlation between age and the About-Me Questionnaire (Identification with the team), r (46) = 0.313, *p* = 0.034. Additionally, checking for group differences between the various education levels did not show any significant differences toward the (modified) About-Me Questionnaire (Identification with the chat partner) as well as the GES (Perception of technology of one's self) and the AGES (Perception of the chat partner as part of one's self). The mean scores of both groups revealed a medium perceived workload. However, to check for a potential difference, we conducted a *t*-test for independent samples due to the non-significant Levene and Shapiro–Wilk tests. Overall, there was no significant difference between the VA group (M = 10.7, SD = 3.16) and the human chat partner group (M = 11.2, SD = 3.65), *p* = 0.611 and *d* = −0.129. Furthermore, the data did not show a difference between the VA chat partner group (M = 2.88, SD = 1.56) and the human chat partner group (M = 3.11, SD = 1.89) regarding the satisfaction score after the chat interaction, *p* = 0.113 and *d* = −0.134.

To check whether satisfaction with the interaction and perceived workload are related, a correlation was

Table 3 Validation of measurements

	Composite Reliability	Cronbach's α	AVE	About-Me	Modified About-Me	GES
About-Me	.780	.759	.477	–	–	–
Modified About-Me	.843	.835	.576	r = −.003	–	–
GES	.840	.839	.471	r = .111	r = .323*	–
AGES	.891	.886	.577	r = .152	r = .589***	r = .467***

Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 4 Model coefficients towards social identification with the chat partner (modified About-Me scale)

Predictor	Estimate	SE	t	*p*
Group: human–VA	0.0606	0.2384	0.254	.801
Age	−0.0147	0.0188	0.254	.439
Gender	0.1421	0.3751	0.379	.707
Satisfaction	−0.1687	0.0774	−2.180	.035*
NASATLX	0.0718	0.0414	1.734	.091

Note: * $p < .05$

Table 5 Model coefficients towards identification with the chat partner as part of one's self (AGES)

Predictor	Estimate	SE	t	*p*
Group: human–VA	0.03410	0.3284	0.104	.918
Age	0.00897	0.0259	0.346	.732
Gender	−0.37018	0.5166	−0.717	.478
Satisfaction	−0.16296	0.1066	−1.529	.317
NASATLX	0.05785	0.0570	1.014	.317

calculated between the two variables. To reveal insights about the two groups, we conducted correlations separately for each group. Satisfaction was positively correlated with perceived workload r (24) = 0.662, $p < 0.001$ in the VA group but not in the human group, r (22) = 0.204, $p = 0.363$. Table 2 presents further significant correlations in the VA group between perceived satisfaction and the single items of the NASA-TLX score.

5 Results

In this section, first, we check the observed major scales' (GES, AGES, About-Me, and Modified About-Me) reliability and validity measures (Cronbach and Meehl 1955; O'Leary-Kelly and Vokurka 1998; Peters 2018). Second, we introduce the results regarding social identity theory and the extended self. Table 3 summarizes the values for composite reliability, average variance extracted (AVE), and construct validity. The comprehensive results are shown in the Appendix (available online via http://link.springer.com), including factor loadings as well as correlation coefficients for each item of the major scales. In summary, the described constructs explain on average more than 50% of the variance (Table 3). Regarding the validity measurements, construct validity shows that the modified About-Me Questionnaire might be linked to the AGES.

5.1 Social Identity

To check for potential group differences regarding the distinct social identity questionnaires, we conducted a one-way ANOVA. According to Levene's test for equality of variances, we cannot assume equality for the collective identity orientation scale (F (1,44) = 6.294, $p = 0.016$), thus we chose the more robust Welch's one-way ANOVA. For collective identity orientation, the VA group (M = 2.18, SD = 0.364) differs significantly from the human (M = 2.82, SD = 0.711) group, F (1,30.7), $p < 0.001$.

To examine social identification with the specific chat partner (bot or human), a linear regression model was calculated that predicts the score on the modified About-Me Questionnaire based on the participant's group and the control variables age, gender, satisfaction, and perceived workload. According to Levene's test of equality of variances ($p = 0.484$) and the Shapiro–Wilk test of normality ($p = 0.713$), we assume equality of variances as well as normal distribution. Results of the multiple linear regression model indicated no significant effect overall, F (5,49) = 1.44, $p = 0.230$, $R^2 = –0.153$. The individual predictors were examined further and indicated that satisfaction (t = –2.18, $p = 0.035$) is a significant predictor in the model (Table 4).

H1 stated that virtual collaboration with a VA, compared to a human partner, affects social identity, that is, the degree of identification with the chat partner. This is not supported by the findings.

To test within each group whether identification with the teams and colleagues differs before and after solving the task, we conducted a paired samples t-test for group differences with a 95% confidence interval and the two measurements of the About-Me Questionnaire as paired variables for each group. For the VA group, the Shapiro–Wilk test of normality was non-significant (p = 0.173), and no violation of normality was therefore assumed. On average in the VA group, the first measurement (M = 3.58, SD = 0.810) of the About-Me Questionnaire was slightly higher than the second measurement (M = 3.34, SD = 0.638). This difference was significant t (23) = 3.15, p = 0.004, with a medium-sized effect (d = 0.64). Therefore, the results support H2, indicating that people who collaborate with VAs indeed identify less with their human team after interaction with the VA than they did before. For the human group, the Shapiro–Wilk test of normality was also non-significant (p = 0.056), so no violation of normality was assumed. Thus, a paired samples t-test was conducted for the human group. The test showed no significant differences (p = 0.773, d = −0.063) between the first measurement of the About-Me Questionnaire (M = 3.38, SD = 0.427) and the second measurement (M = 3.33, SD = 0.633).

5.2 The Extended Self

To examine the role of the extended self in the context of social identity and virtual collaboration, we conducted group comparisons and correlations. We analyzed the score of the GES as well as the score of the AGES regarding the chat interaction used in the experiment.

To reveal potential influences of the groups and control variables on the identification with the chat partner (AGES) as part of one's self, we applied a linear regression model. Levene's test for equality of variances was not significant for the AGES (p = 0.279); thus, equality of variances was assumed. Results of the multiple linear regression model indicated no significant effect of the group (human or VA) or the control variables age, gender, satisfaction, and perceived workload on the identification with the chat-partner as part of one's self (AGES), F (5,49) = 0.666, p = 0.652, R^2 = −0.0768. The individual predictors were examined further, and none of them were significant (Table 5). These results do not support an impact of the groups, thus H3 is not supported by the findings.

Furthermore, we investigated the relationship between the two scales of the extended self and the perception of the chat partner (VA and human) as being part of one's social group at work. To this end, we conducted a bivariate correlation overall for both groups as well as separately for each group. Overall, the GES score, r (46) = 0.467, p = 0.001, and AGES score, r (46) = 0.589, $p < 0.001$, showed significant positive correlations with the modified About-Me Questionnaire. Analyzing the relationship for the VA group revealed a significant positive correlation between the GES score and the modified About-Me Questionnaire, r (24) = 0.486, p = 0.016. Likewise, the AGES score correlates significantly, r (24) = 0.641, $p < 0.001$. The human chat partner group showed only a significantly positive correlation for the AGES score and the modified About-Me Questionnaire, p = 0.009, r (22) = 0.540. Therefore, the correlation between the GES score and the modified About-Me Questionnaire was not significant, p = 0.336, r = 0.215. To summarize, the results do not support a negative relationship between individuals' identification with the team and individuals' identification with technology as a part of their extended self (H4). However, the results revealed a positive relationship.

6 Discussion

6.1 Key Findings

In this study, we examined how a VA affects social identity and the extended self in virtual collaboration. First, we did not find a significant impact of virtual collaboration with a VA, compared to a human partner, on social identity, that is, on the degree of identification with the team (H1). In this context, VAs may do not differ as a team member compared to a human. This is consistent with the results of Edwards et al. (2019), who found that VAs could act as equal social actors.

However, a key finding of this paper is that people who collaborate with VAs identify less with their (human) team after their interaction with the VA than they did before (H2). This medium-sized effect indicates that working with VAs could influence the social identity of a person in the context of virtual collaboration. This may be explained by the fact that the person feels more independent and able to solve the task alone. Even if, according to Young-Jae et al. (2020), people increasingly face difficulties in expressing the uniqueness of humans compared to AI applications, VAs seem to reduce the social identification with team members. This may be explained by the feeling that people experience less connection to their team after interacting with the VA solely. However, this does not appear to be due to an emotional attachment to the VA as You and Robert (2018) found a connection between team identity and emotional attachment to VAs. Therefore, further questions arise for future IS research: How should we design a VA in order to strengthen the feeling of being connected to the team? How important is the role of identification with one's own team for future work? What impact will VAs have on team collaboration? What implications will VAs have on the digital workplace?

There is no significant difference in the perceived workload of the task and the achieved score between the group supported by a VA and the group assisted by another human. The workload of solving the CPM assisted by the VA is therefore neither perceived as higher nor as lower. This result is contrary to Moreno et al. (2001) and Brachten et al. (2020), who were able to show that individuals supported by VAs outperform humans who did not use a VA. Furthermore, Mechling et al. (2010) demonstrated that groups advised by a VA reach better outcomes. However, a positive lesson that can be drawn from this is that the task-solving with the VA did not put any additional strain on the participants in solving the tasks. In this respect, the support by a VA seems to be similar to the support by another person.

The results do not suggest an influence of collaboration with a VA or a human chat partner on the perception of the respective collaboration partner as part of one's extended self (H3). According to identity research, the formation of identity and its extension is a dynamic process that adapts over time (Burke and Stets 2009; Carter et al. 2015). At the point of introducing a new technology, the participants did not perceive the VA and the human chat partner differently regarding the chat partner as a resource for maintaining or enhancing the self.

6.2 Implications for Theory: The New Concept of Virtually Extended Identification

As a key finding and in contradiction to H4, the study revealed that someone who identifies with their team members is also more likely to identify with the technology as a part of their extended self and vice versa. This highlights a possible connection between the theory of social identity and the concept of the extended self, as some literature hinted at. We found a positive correlation between the individual's identification with the team and the individual's identification with technology as a part of their extended self (H4 not supported). Particularly, for social identification with technology, such as VAs as team members (Seeber et al. 2020a), the underlying concept of the extended self could be considered to explain upcoming interactions. Considering individuals' mental processes in social groups, individuals divide other team members into either their in-group or out-group. They apply social rules and determine the value of their own group related to other groups (Tajfel and Turner 1986). This conceptualization does not sufficiently consider that technology, specifically, a VA, is capable of being a virtual team member. Working with a VA as a virtual team member might enrich one's social group by perceiving the VA as a team member of the group (external perspective). Furthermore, a VA might support one's self-esteem by positively identifying with the VA's characteristics and capabilities, which might lead to enhancing one's human capabilities (internal perspective). Therefore, VAs may be externally attributed to one's in-group as a team member or be part of one's in-group by internally attributing the VA to one's self. However, past research does not differentiate the two pathways that we examined with H4.

People use newly introduced technology such as a VA and identify with the capabilities and characteristics of these supportive tools when they start to compare themselves with the VA. On one hand, people feel connected to this technology that might lead to improving their own capabilities with the aid of a VA. On the other hand, people then perceive the VA as a social team mate, according to Seeber et al. (2018). This can also be the other way around. Therefore, both concepts are necessary to understand how human behavior is influenced by newly introduced technology such as VAs. Furthermore, analyzing the construct validity has shown that the constructs of the extended self and social identity theory directed toward the VA are connected (Cronbach and Meehl 1955; O'Leary-Kelly and Vokurka 1998). We hence derive that for the context of virtual collaboration, the construct *identification with team (members)* of the social identity theory and the concept of the *extended self* are intertwined. Each may represent different facets of the same underlying construct. This becomes evident regarding the aspects of social comparison and positive distinctiveness of the social identity theory and the process of extending the self. People consider personal attributes, other people, groups (e.g., values of the group), or abstract ideas (e.g., morals of society) in regard to their self when forming the self. An extension of the self can take place by regarding these (social) aspects through control (e.g., a technology), knowledge (e.g., a person), or a feeling of belonging (Tajfel and Turner 1986; Belk 1988; Carter et al. 2015). Thus, people compare themselves with people and technology to determine and extend their own identity. This also happens with possessions, such as technology at the workplace (Tian and Belk 2005). By positively identifying with the VA, positive distinctiveness can be brought about, especially in the workplace.

Our findings suggest a positive connection between social identity theory and extended self (H4). We therefore propose combining these two aspects of identification into the overarching construct of *virtually extended identification* to understand the relationships evolving in virtual collaboration with VAs (see Fig. 3). Virtually extended identification describes the process of maintaining and extending the self by comparing the current self with a VA.

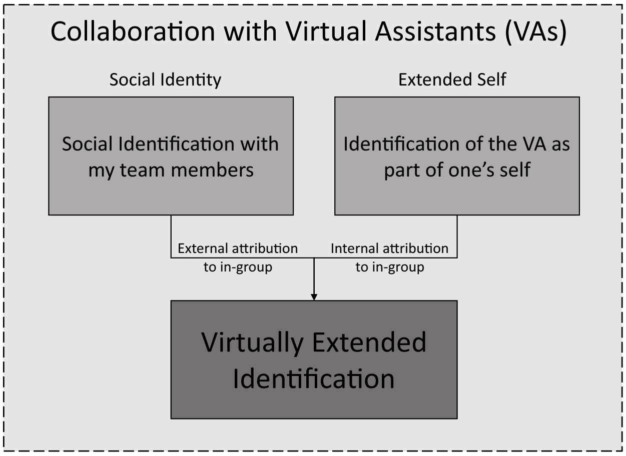

Collaboration with Virtual Assistants (VAs)
Social Identity
Extended Self
Social Identification with my team members
Identification of the VA as part of one's self
External attribution to in-group
Internal attribution to in-group
Virtually Extended Identification

Fig. 3 Symbolic formation of social identity and the extended self in the context of virtual collaboration with technology such as VAs

On the one hand, the VA substitutes the role of a human collaborator, according to Seeber et al. (2018), Demir et al. (2020), and Panganiban et al. (2020). On the other hand, the VA is also considered as technology, according to Schwabe (2003), Bajwa et al. (2007), Frohberg and Schwabe (2006), and Vahtera et al. (2017). Thus, the observed relationship between the extended self and the social identification with the VA reveals that a VA as a supportive conversational technology has a dual function. This means that people can assess a VA as a social actor as well as a form of technology at the same time. Therefore, virtually extended identification describes the degree to which a person's identity matches the perceived identity of the VA as a team member (social actor as an external attribution to the in-group) and the degree to which the capabilities of the VA are attributed to the person's self (internal attribution to the in-group by the identification of the VA's characteristics, values, and capabilities with the self). This dual function of the VA is also based on the results suggesting that VAs do not significantly differ compared to a human chat partner regarding influence on perceived workload, performance (H1 and H3 not supported). However, satisfaction might have an impact on the identification with the chat partner in the context of virtual collaboration as the findings imply. Thus, companies could save valuable resources by deploying VAs in virtual collaboration as a chat partner. VAs should be deployed as both supportive tools to assist work-related tasks and as members of virtual teams to increase social identity and positive distinctiveness. In this way, the positive aspects of both theories (Lin 2015; Vahtera et al. 2017) could be used to achieve an overarching goal more efficiently. The creation of a social presence through social cues (Feine et al. 2019) could further reinforce these aspects (Franceschi et al. 2009).

Thus, one of the most relevant findings of this study is that social identity and the extended self in virtual collaboration with VAs are not contradictory, as assumed in H4. VAs can be perceived simultaneously as team members and as tools. The boundaries between technology as a collaboration platform and tool and technology as a partner for virtual collaboration seem to blur. However, the question arises as to whether our findings can be generalized since we examined a specific VA in our experiment. In this respect, recent research is currently using many VAs, chatbots, and conversational agents that are purely text-based agents (Hofeditz et al. 2019, p. 201; Diederich et al. 2020; Brachten et al. 2020). We used the social cues that are effective according to current knowledge (Feine et al. 2019) and tried to keep the interference factors, such as the influence of a time limit on team performance (Massey et al. 2003), as low as possible. Our insight into the relationship between social identity theory and the extended self in the context of virtual collaboration with VAs leads to an advanced understanding of machines as teammates and can be explained by the existing IS literature (Schwabe 2003; Waizenegger et al. 2020; Seeber et al. 2020a, b).

6.3 Limitations and Further Research

This study examined the effects of a newly introduced technology. It may be possible that the perception of the VA changes over time by using the VA for a longer period. Further studies may use and compare these findings with studies where VAs are used over longer periods of time. The level of anthropomorphism of a VA and the use of different social cues might also influence the perception of a VA. This aspect should be considered in future research.

As we focused on understanding the perception of VAs in the context of social identity and extended self, we examined one cultural background which is Central European. Further studies may consider cross-cultural differences in regard to VA adoption. Moreover, further studies may aim for a larger sample size to show possible unrevealed effects. Furthermore, we strongly recommend testing the proposed construct of virtual identification in different collaborative scenarios to take the next steps in understanding identification in the context of virtual collaboration.

Moreover, not only text-based communication but also interaction via speech may have an influence on the perception of VAs (Edwards et al. 2019). Additionally, the collaboration platform used in which the VA was integrated could also have influenced the social identity (Hu et al. 2017). Furthermore, the virtual collaboration environment might also be an influencing factor on the perception of the VA. We suggest that future research consider potential differences in virtual collaboration between distinct environments.

7 Conclusion

This study provides new insights regarding social identity theory as well as the concept of the extended self in the context of virtual collaboration. First, it was shown that people who work with VAs identify less with their (human) team after their interaction with a VA. Therefore, collaborative VAs may influence the social identity of a person. Second, this study highlights that someone who identifies the VA as part of their extended self is also more likely to identify with (virtual) team members and vice versa. The revealed intertwining emphasized that research needs to change its understanding of (social) identification in the context of virtual collaboration with VAs. Neither concept should be regarded in isolation.

This study contributes to social identity theory as well as the extended self by proposing a new construct to understand identification with team members and technology in a collaborative context. The study reveals that the relationship between social identification with (virtual) team members and expanding the self through technology such as VAs is not contradictory but rather that they complement each other. VAs are not only perceived as resources to maintain and extend one's identity but also as social actors. This implies that research should not separate these concepts but rather combine their specific aspects to understand human behavior in virtual collaboration. To this end, items of both constructs may be combined and evaluated to develop the new virtually extended identification construct. This concept may be better suited for understanding human behavior in the changing landscape of virtual collaboration.

This study also provides practical contributions. VAs are a collaborative tool with a low entry barrier. The findings suggest that the support of a VA is similar to that of a human. Thus, organizations could save valuable resources by using VAs to support employees in their tasks. Especially in the context of a newly introduced technology, one could expect the effort needed to learn the technology to lead to an increase in perceived workload, but no significant effect was observed. However, the results indicate that the collaboration with a VA might lower the identification with other team members. As a worst-case scenario, employees do not feel part of the human team in return. Thus, decision makers should take measures to encourage the continued identification with other colleagues when introducing such technology within the organization. However, people might identify VAs as resources for expanding their own capabilities, but at the same time VAs might be seen as social actors during collaboration. Overall, VAs are a resource-saving tool that managers may use to support their human employees. In this context, the introduction of VAs should be accompanied by measures to support the continued social identification with other colleagues, such as social events or gatherings.

Funding Open Access funding enabled and organized by Projekt DEAL. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska- Curie grant agreement No 823866.

References

Andres HP, Shipps BP (2019) Team learning in technology-mediated distributed teams. J Inf Syst Educ 21:10

Araujo T (2018) Living up to the chatbot hype: the influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Comput Hum Behav 85:183–189. https://doi.org/10.1016/j.chb.2018.03.051

Aron A, Aron E, Norman C (2003) Self-expansion motivation and including other in the self. In: Fletcher GJO, Clark MS (eds) Blackwell handbook of social psychology: interpersonal processes. Blackwell, Oxford

Bajwa D, Lewis L, Pervan G et al (2007) Organizational assimilation of collaborative information technologies: global comparisons. In: 2007 40th annual Hawaii international conference on system sciences

Bartels J, van Vuuren M, Ouwerkerk JW (2019) My colleagues are my friends: the role of facebook contacts in employee identification. Manag Commun Q 33:307–328. https://doi.org/10.1177/0893318919837944

Belk RW (1988) Possessions and the extended self. J Consum Res 15:139. https://doi.org/10.1086/209154

Belk RW (2013) Extended self in a digital world. J Consum Res 40:477–500. https://doi.org/10.1086/671052

Benbya H, Leidner D (2018) How Allianz UK used an idea management platform to harness employee innovation. MIS Q Exec 17:900–933

Berg MM (2015) NADIA: a simplified approach towards the development of natural dialogue systems. In: Biemann C, Handschuh S, Freitas A et al (eds) Natural language processing and information systems. Springer, Cham, pp 144–150

Bittner E, Oeste-Reiß S, Leimeister JM (2019) Where is the bot in our team? Toward a taxonomy of design option combinations for conversational agents in collaborative work. In: Proceedings of the 52nd Hawaii international conference on system sciences

Brachten F, Brünker F, Frick NRJ et al (2020) On the ability of virtual agents to decrease cognitive load: an experimental study. Inf Syst E-Bus Manag 18:187–207. https://doi.org/10.1007/s10257-020-00471-7

Brown R (2000) Social identity theory: past achievements, current problems and future challenges. Eur J Soc Psychol 30:745–778. https://doi.org/10.1002/1099-0992(200011/12)30:6%3c745::AID-EJSP24%3e3.0.CO;2-O

Burke P (2006) Identity change. Soc Psychol Q 69:81–96

Burke PJ, Stets JE (2009) Identity theory. Oxford University Press, New York

Canonico M, Russis LD (2018) A comparison and critique of natural language understanding tools. In: The Ninth International Conference on Cloud Computing, GRIDs, and Virtualization

Cao A, Chintamani KK, Pandya AK, Ellis RD (2009) NASA TLX: software for assessing subjective mental workload. Behav Res Methods 41:113–117. https://doi.org/10.3758/BRM.41.1.113

Carter M, Grover V (2015) Me, my self, and I(T): conceptualizing information technology identity and its implications. MIS Q 39:931–957. https://doi.org/10.25300/MISQ/2015/39.4.9

Carter M, Grover V, Clemson University (2015) Me, my self, and I(T): conceptualizing information technology identity and its implications. MIS Q 39:931–957. https://doi.org/10.25300/MISQ/2015/39.4.9

Carter M, Grover V, Thatcher JB (2012) Mobile devices and the self: developing the concept of mobile phone identity. In: Lee I (ed) Strategy, adoption, and competitive advantage of mobile services in the global economy. IGI Global, Hershey

Changizi A, Lanz M (2019) The comfort zone concept in a human–robot cooperative task. In: Ratchev S (ed) Precision assembly in the digital age. Springer, Cham, pp 82–91

Chung H, Iorga M, Voas J, Lee S (2017) Alexa, can i trust you? Comput 50:100–104. https://doi.org/10.1109/MC.2017.3571053

Clayton RB, Leshner G, Almond A (2015) The extended iSelf: the impact of iphone separation on cognition, emotion, and physiology. J Comput-Mediat Commun 20:119–135. https://doi.org/10.1111/jcc4.12109

Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52:281–302. https://doi.org/10.1037/h0040957

Dahling JJ, Gutworth MB (2017) Loyal rebels? A test of the normative conflict model of constructive deviance. J Organ Behav 38:1167–1182. https://doi.org/10.1002/job.2194

Davenport T (2018) The AI advantage how to put the artificial intelligence revolution to work, 1st edn. MIT Press, Cambridge

de Vreede G-J, Briggs RO (2005) Collaboration engineering: designing repeatable processes for high-value collaborative tasks. In: Proceedings of the 38th Hawaii International Conference on System Sciences – 2005

Demir M, McNeese NJ, Cooke NJ (2020) Understanding human–robot teams in light of all-human teams: aspects of team interaction and shared cognition. Int J Hum Comput Stud 140:102436. https://doi.org/10.1016/j.ijhcs.2020.102436

Dias M, Pan S, Tim Y (2019) Knowledge embodiment of human and machine interactions: robotic-process-automation at the Finland government. In: Proceedings of the 27th European Conference on Information Systems

Diederich S, Brendel AB, Kolbe LM (2020) Designing anthropomorphic enterprise conversational agents. Bus Inf Syst Eng 62:193–209. https://doi.org/10.1007/s12599-020-00639-y

Dittmar H (2011) Material and consumer identities. In: Schwartz SJ, Luyckx K, Vignoles VL (eds) Handbook of identity theory and research. Springer, New York, pp 745–769

Edwards C, Edwards A, Stoll B et al (2019) Evaluations of an artificial intelligence instructor's voice: social Identity Theory in human–robot interactions. Comput Hum Behav 3:357–362. https://doi.org/10.1016/j.chb.2018.08.027

Ellemers N (2004) Motivating individuals and groups at work: a social identity perspective on leadership and group performance. Acad Manage Rev 29:459–478

Feine J, Gnewuch U, Morana S, Maedche A (2019) A taxonomy of social cues for conversational agents. Int J Hum Comput Stud 132:138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

Franceschi K, Lee RM, Zanakis SH, Hinds D (2009) Engaging group e-learning in virtual worlds. J Manag Inf Syst 26:73–100. https://doi.org/10.2753/MIS0742-1222260104

Frohberg D, Schwabe G (2006) Skills and motivation in ad-hoc-collaboration. Collect Collab Electron Commer Technol Res. https://doi.org/10.5167/uzh-61366

Galy E, Cariou M, Mélan C (2012) What is the relationship between mental workload factors and cognitive load types? Int J Psychophysiol 83:269–275. https://doi.org/10.1016/j.ijpsycho.2011.09.023

Gnewuch U, Morana S, Maedche A (2017) Towards designing cooperative and social conversational agents for customer service. In: International conference on information systems, p 15

Gnewuch U, Yu M, Maedche A (2020) The effect of perceived similarity in dominance on customer self-disclosure to chatbots in conversational commerce. In: 28th European conference on information systems

Goodbody J (2005) Critical success factors for global virtual teams. Strateg Commun Manag 9:18–21

Guegan J, Segonds F, Barré J et al (2017) Social identity cues to improve creativity and identification in face-to-face and virtual groups. Comput Hum Behav 77:140–147. https://doi.org/10.1016/j.chb.2017.08.043

Hart SG (2006) Nasa-task load index (NASA-TLX); 20 years later. In: Proceedings of the human factors and ergonomics society 50th annual meeting – 2006, p 5

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock PA, Meshkati N (eds) Advances in psychology. Elsevier, North-Holland, pp 139–183

Haslam A (2004) Psychology in organizations: the social identity approach, 2nd edn. SAGE, London

Hassell MD, Cotton JL (2017) Some things are better left unseen: toward more effective communication and team performance in video-mediated interactions. Comput Hum Behav 73:200–208. https://doi.org/10.1016/j.chb.2017.03.039

Hofeditz L, Ehnis C, Bunker D et al (2019) Meaningful use of social bots? possible applications in crisis communication during disasters. In: Proceedings of the 27th European conference on information systems

Hu M, Zhang M, Wang Y (2017) Why do audiences choose to keep watching on live video streaming platforms? An explanation of dual identification framework. Comput Hum Behav 75:594–606. https://doi.org/10.1016/j.chb.2017.06.006

Kenny EJ, Briner RB (2013) Increases in salience of ethnic identity at work: the roles of ethnic assignation and ethnic identification. Hum Relat 66:725–748. https://doi.org/10.1177/0018726712464075

Klimchak M, Ward A-K, Matthews M et al (2019) When does what other people think matter? The influence of age on the motivators of organizational identification. J Bus Psychol 34:879–891. https://doi.org/10.1007/s10869-018-9601-6

Knijnenburg BP, Willemsen MC (2016) Inferring capabilities of intelligent agents from their external traits. ACM Trans Interact Intell Syst 6:1–25. https://doi.org/10.1145/2963106

Knote R, Janson A, Söllner M, Leimeister JM (2019) Classifying smart personal assistants: an empirical cluster analysis. In: Proceedings of the 52nd Hawaii international conference on system sciences

Kohler F, Matzler, et al (2011) Co-creation in virtual worlds: the design of the user experience. MIS Q 35:773. https://doi.org/10.2307/23042808

Kozlowski S, Ilgen D (2007) The science of team success. Sci Am Mind 18:54–61. https://doi.org/10.1038/scientificamericanmind0607-54

Lamontagne L, Laviolette F, Khoury R, Bergeron-Guyard A (2014) A framework for building adaptive intelligent virtual assistants. In: Software engineering/811: parallel and distributed computing and networks/816: artificial intelligence and applications. ACTAPRESS, Innsbruck, Austria

Leary MR, Tangney JP (2011) Handbook of self and identity, 2nd edn. Guilford Press, New York

Lee C, Jung S, Kim S, Lee GG (2009) Example-based dialog modeling for practical multi-domain dialog system. Speech Commun 51:466–484. https://doi.org/10.1016/j.specom.2009.01.008

Lin C-P (2015) Predicating team performance in technology industry: theoretical aspects of social identity and self-regulation. Technol Forecast Soc Change 98:13–23. https://doi.org/10.1016/j.techfore.2015.05.017

Luger E, Sellen A (2016) "Like having a really bad pa": the gulf between user expectation and experience of conversational agents. In: Proceedings of the 2016 CHI conference on human factors in computing systems

Maedche A, Legner C, Benlian A et al (2019) AI-based digital assistants: opportunities, threats, and research perspectives. Bus Inf Syst Eng 61:535–544. https://doi.org/10.1007/s12599-019-00600-8

Maniscalco U, Messina A, Storniolo P (2020) The human–robot interaction in robot-aided medical care. In: Zimmermann A, Howlett RJ, Jain LC (eds) Smart innovation, systems and technologies. Springer, Split, pp 233–242

Maras P, Thompson T, Gridley N, Moon A (2018) The "about me" questionnaire: factorial structure and measurement invariance. J Psychoeduc Assess 36:379–391. https://doi.org/10.1177/0734282916682909

Massey AP, Montoya-Weiss MM, Hung Y-T (2003) Because time matters: temporal coordination in global virtual project teams. J Manag Inf Syst 19:129–155. https://doi.org/10.1080/07421222.2003.11045742

McDuff D, Czerwinski M (2018) Designing emotionally sentient agents. Commun ACM 61:4–83. https://doi.org/10.1145/3186591

McTear MF (2017) The rise of the conversational interface: a new kid on the block? In: Future and emerging trends in language technology. Machine learning and big data. Springer, Seville

Mechling LC, Gast DL, Seid NH (2010) Evaluation of a personal digital assistant as a self-prompting device for increasing multistep task completion by students with moderate intellectual disabilities. Educ Train Autism Dev Disabil 45:422–439

Moreno R, Mayer RE, Spires HA, Lester JC (2001) The case for social agency in computer-based teaching: do students learn more deeply when they interact with animated pedagogical agents? Cogn Instr 19:177–213. https://doi.org/10.1207/S1532690XCI1902_02

Morrissey K, Kirakowski J (2013) 'Realness' in chatbots: establishing quantifiable criteria. In: Kurosu M (ed) Human–computer interaction. Interaction modalities and techniques. Springer, Berlin, pp 87–96

Mueller SK, Mendling J, Bernroider EWN (2019) The roles of social identity and dynamic salient group formations for ERP program management success in a postmerger context. Inf Syst J 29:609–640. https://doi.org/10.1111/isj.12223

Nasirian F, Ahmadian M (2017) AI-based voice assistant systems: evaluating from the interaction and trust perspectives. In: Americas conference on information systems. p 10

Norman D (2017) Design, business models, and human-technology teamwork: as automation and artificial intelligence technologies develop, we need to think less about human–machine interfaces and more about human–machine teamwork. Res Technol Manag 60:26–30. https://doi.org/10.1080/08956308.2017.1255051

Noyes JM, Bruneau DPJ (2007) A self-analysis of the NASA-TLX workload measure. Ergonomics 50:514–519. https://doi.org/10.1080/00140130701235232

Nunamaker JF, Derrick DC, Elkins AC et al (2011) Embodied conversational agent-based kiosk for automated interviewing. J Manag Inf Syst 28:17–48. https://doi.org/10.2307/41304605

O'Leary-Kelly SW, Vokurka RJ (1998) The empirical assessment of construct validity. J Oper Manag 16:387–405. https://doi.org/10.1016/S0272-6963(98)00020-5

Panganiban AR, Matthews G, Long MD (2020) Transparency in autonomous teammates: intention to support as teaming information. J Cogn Eng Decis Mak 14:174–190. https://doi.org/10.1177/1555343419881563

Pepple DG, Davies EMM (2019) Co-worker social support and organisational identification: does ethnic self-identification matter? J Manag Psychol 34:573–586. https://doi.org/10.1108/JMP-04-2019-0232

Peters G-JY (2018) The alpha and the omega of scale reliability and validity: why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. Eur Health Psychol 16:56–59. https://doi.org/10.31234/osf.io/h47fv

Pickard MD, Burns MB, Moffitt KC (2013) A theoretical justification for using embodied conversational agents (ECAs) to augment accounting-related interviews. J Inf Syst 27:159–176. https://doi.org/10.2308/isys-50561

Plotnick L, Hiltz SR, Privman R (2016) Ingroup dynamics and perceived effectiveness of partially distributed teams. IEEE Trans Prof Commun 59:203–229. https://doi.org/10.1109/TPC.2016.2583258

Porck JP, Matta FK, Hollenbeck JR et al (2019) Social identification in multiteam systems: the role of depletion and task complexity. Acad Manage J 62:1137–1162. https://doi.org/10.5465/amj.2017.0466

Quarteroni S (2018) Natural language processing for industry: ELCA's experience. Inform Spektrum 41:105–112. https://doi.org/10.1007/s00287-018-1094-1

Rubio S, Díaz E, Martín J, Puente JM (2004) Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. Appl Psychol 53:61–86. https://doi.org/10.1111/j.1464-0597.2004.00161.x

Russel S, Norvig P (2016) Artificial intelligence: a modern approach. Addison Wesley, Munich

Schuetzler RM, Giboney JS, Grimes GM, Nunamaker JF (2018) The influence of conversational agent embodiment and conversational relevance on socially desirable responding. Decis Support Syst 114:94–102. https://doi.org/10.1016/j.dss.2018.08.011

Schultze U (2010) Embodiment and presence in virtual worlds: a review. J Inf Technol 25:434–449. https://doi.org/10.1057/jit.2010.25

Schwabe G (2003) Growing an application from collaboration to management support: the example of Cupark. doi:10/gg4ms7

Scott C, Lundstrom WJ (1990) Dimensions of possession satisfactions: a preliminary analysis. J Satisf Dissatisfaction Complain Behav 3:100–1004

Seeber I, Bittner E, Briggs RO, et al (2018) Machines as teammates: a collaboration research agenda. In: Proceedings of the 51st Hawaii international conference on system sciences

Seeber I, Bittner E, Briggs RO et al (2020a) Machines as teammates: a research agenda on AI in team collaboration. Inf Manage 57:103174. https://doi.org/10.1016/j.im.2019.103174

Seeber I, Waizenegger L, Seidel S et al (2020b) Collaborating with technology-based autonomous agents: issues and research opportunities. Internet Res 30:1–18. https://doi.org/10.2139/ssrn.3504587

Shamekhi A, Liao QV, Wang D, et al (2018) Face value? Exploring the effects of embodiment for a group facilitation agent. In: Conference on human factors in computing systems

Sivadas E, Machleit KA (1994) A scale to determine the extent of object incorporation in the extended self. Mark Theory Appl 5:143–149

Spohrer J, Banavar G (2015) Cognition as a service: an industry perspective. AI Mag 36:71–86. https://doi.org/10.1609/aimag.v36i4.2618

Stets JE, Biga CF (2003) Bringing identity theory into environmental sociology. Soc Theory 21:398–423. https://doi.org/10.1046/j.1467-9558.2003.00196.x

Stets JE, Burke PJ (2000) Identity theory and social identity theory. Soc Psychol Q 63:224. https://doi.org/10.2307/2695870

Stieglitz S, Brachten F, Kissmer T (2018) Defining bots in an enterprise context. In: International conference on information systems

Tajfel H, Turner JC (1986) The social identity theory of intergroup behavior. In: Austin WG, Worchel S (eds) Psychology of intergroup relation. Hall Publishers, Chicago, pp 7–24

Tian K, Belk RW (2005) Extended self and possessions in the workplace. J Consum Res 32:297–310. https://doi.org/10.1086/432239

Vahtera P, Buckley PJ, Aliyev M et al (2017) Influence of social identity on negative perceptions in global virtual teams. J Int Manag 23:367–381. https://doi.org/10.1016/j.intman.2017.04.002

Van Ostrand A, Wolfe S, Arredondo A et al (2016) Creating virtual communities that work: best practices for users and developers of e-collaboration software. Int J E-Collab 12:41–60. https://doi.org/10.4018/IJeC.2016100104

Verhagen T, van Nes J, Feldberg F, van Dolen W (2014) Virtual customer service agents: using social presence and personalization to shape online service encounters. J Comput Mediat Commun 19:529–545. https://doi.org/10.1111/jcc4.12066

Vignoles VL, Schwartz SJ, Luyckx K (2011) Introduction: toward an integrative view of identity. In: Schwartz SJ, Luyckx K, Vignoles VL (eds) Handbook of identity theory and research. Springer, New York, NY, pp 1–27

Waizenegger L, Seeber I, Dawson G, Desouza KC (2020) Conversational agents – exploring generative mechanisms and second-hand effects of actualized technology affordances. In: Proceedings of the 53rd Hawaii international conference on system sciences

Wang W (2017) Smartphones as social actors? social dispositional factors in assessing anthropomorphism. Comput Hum Behav 68:334–344. https://doi.org/10.1016/j.chb.2016.11.022

Wang W, Siau K (2018) Artificial intelligence: a study on governance, policies, and regulations. In: MWAIS 2018 proceedings

Yan JK, Leidner DE, Benbya H (2018) Differential innovativeness outcomes of user and employee participation in an online user innovation community. J Manag Inf Syst 35:900–933. https://doi.org/10.1080/07421222.2018.1481669

You S, Robert LP Jr (2018) Emotional attachment, performance, and viability in teams collaborating with embodied physical action (EPA) robots. J Assoc Inf Syst 19:377–407. https://doi.org/10.17705/1jais.00496

Young-Jae C, Baek S, Ahn G et al (2020) Compensating for the loss of human distinctiveness: the use of social creativity under human–machine comparisons. Comput Hum Behav 103:80–90. https://doi.org/10.1016/j.chb.2019.08.027

Paper 6: Mind Attribution is Key to Understanding Virtual Influencer Perception

Type (Ranking, Impact Factor)	Journal article (A, 7.79)
Status	Major Revisions (2nd round)
Rights and permissions	Closed Access
Authors	Hofeditz, L., Nissen, A., Schütte, R., Mirbabaie, M., Stieglitz, S.
Year	Forthcoming
Outlet	Journal of the Association for Information Systems
Permalink / DOI	N/A (under review)
Full citation	Hofeditz, L., Nissen, Anika, Schütte, R., Mirbabaie, M., & Stieglitz, S. (forthcoming). Mind Attribution is Key to Understanding Virtual Influencer Perception. *Journal of the Association for Information Systems.* (under review).

Mind Attribution is Key to Understanding Virtual Influencer Perception

Lennart Hofeditz[1], Anika Nissen[2], Reinhard Schütte[3], Milad Mirbabaie[4] and Stefan Stieglitz[1]

[1]Faculty of Economics and Social Sciences, University of Potsdam
[2]Faculty of Business Administration and Economics, University of Hagen
[3]Faculty of Business Administration and Economics, University of Duisburg-Essen
[4]Faculty of Information Systems and Applied Computer Sciences, University of Bamberg

Abstract

With a simulated human-like identity and a high number of followers, virtual influencers have become a popular phenomenon with growing importance for organizations to promote their brands and products. In contrast to their human counterparts, virtual influencers are easier to control, easy to scale, location-independent, and time-independent. Despite their high potential for marketing campaigns, the knowledge of the key mechanisms in their perception by social media users is ambiguous and raises many questions such as which variables might facilitate positive perceptions by social media users. Across three independent studies, we examined how mind attribution, uncanniness, social presence, and trust explain people's perceptions of these influencers. Our results show that although people cannot always determine whether an influencer is human or not, perceptions of virtual influencers are more negative than for human influencers. We identified that the level of mind attribution to such influencers is key to the perception of the influencer. It becomes evident, that mind attributing processes of influencers take place in brain areas of the mPFC and seem to be encoded with their value for the self. This coding can be seen in self-report perceptions of mind attribution, as well as uncanniness, social presence, and trust. Disclosing virtual influencers as non-human decreased levels of mind attribution, resulting in higher uncanniness.

Keywords— Virtual Influencers, Mind Attribution, Trust, Uncanniness, Social Presence, Social Media, NeuroIS, fNIRS

1 Introduction

Virtual influencers are computer-generated avatars who do not exist in real life, but achieve a large number of followers through regular social media posts (Arsenyan & Mirowska, 2021; Batista & Chimenti, 2021; Moustakas et al., 2020). They become increasingly popular and offer new opportunities for organizations (Choudhry et al., 2022). Virtual influencers promote products of various brands, such as Prada, Samsung, Spotify, or the like, just like real human influencers (Moustakas et al., 2020). Unlike their human counterparts, they minimize risks for companies by offering a high level of controllability and scalability (Batista & Chimenti, 2021). Virtual influencers are designed to achieve a high level of user engagement by gaining users' trust through attractive and authentic content (Batista & Chimenti, 2021; D. Y. Kim & Kim, 2021). Gaining users' trust is one of the central drivers of success in social media (Song & Lee, 2016). Trust generally involves the deliberate decision to engage in a state of mind, establishing a connection and collaboration between two individuals, where one anticipates the utmost from the other, despite harboring uncertainties (Dunn, 2000). Trust in influencers is based on the perceived reliability and integrity and "assures followers that their relationship with the influencer will affect them positively" (D. Y. Kim & Kim, 2021).

One fundamental process in people's decision whether or not to trust someone is the attribution of mind to an interaction partner (ie., by estimating intentions, emotions, and cognitive abilities) (DiYanni et al., 2012; Riedl et al., 2014a). The mind attribution process is explained through the theory of mind (ToM), suggesting an evolutionary-based human ability to attribute a mind to an individual (Mou et al., 2020). Mind attribution can base on an automated and implicit social-perceptual system, and a slower and explicit reflexive-cognitive system (Meinhardt-Injac et al., 2018). Riedl et al. (2014a) concluded that the process of mind attribution works better for human faces than for computer-generated avatars which might result in lower trust towards avatars. However, they based their findings on the perception of human-like, but non photo-realistic avatars. In contrast, photo-realistic human-like virtual influencers are characterized by a high degree of realism (Batista & Chimenti, 2021; D. Y. Kim & Kim, 2021). Through this, they can generate attractiveness which can mitigate their lack of authenticity compared to human influencers (Batista & Chimenti, 2021). This can even result in people falsely believing virtual influencers to be human influencers (Arsenyan & Mirowska, 2021). As a consequence, people might trust virtual influencers even more than human influencers (Nightingale & Farid, 2022) while thinking that they are individual human actors and not computer-generated entities.

One highly relevant factor that increases trust toward other actors is perceived social presence (Hess et al., 2009). Social presence involves the perceived warmth and closeness towards a human or a computer system (Lu et al., 2016). While trust and social presence generally increase with human-likeness levels, certain levels of high but not perfect human-likeness can result in perceived uncanniness which can have negative effects on trust decisions (Bente et al., 2008; Nissen & Jahn, 2021). Uncanniness refers to the unsettling feeling of encountering something that is familiar yet seems strangely unfamiliar (such as a virtual influencer), often triggering discomfort or unease due to a perceived divergence from expected norms or patterns (Dabiran et al., 2022). The term uncanniness refers to the uncanny valley which specifically pertains to the discomfort and eeriness experienced when human-like robots or animated characters closely resemble humans but exhibit subtle discrepancies, which can provoke negative emotional responses (Dabiran et al., 2022). However, people tend to implicitly trust less realistic looking avatars less than real human faces which is signified by both brain activity and behavioral observations (Riedl et al., 2014a).

As a result, there seems to be a positive correlation between trust and mind attribution (DiYanni et al., 2012), and these evaluations can be seen in distinct brain activation (Mustafa et al., 2017; Riedl et al., 2014a; Winston et al., 2002). Thereby, adding brain activation measurement has the great advantage of also uncovering implicit processing of virtual influencers within seconds of looking at an influencer's post. To get a more holistic understanding of virtual influencer perception, it is therefore fruitful to include both brain imaging and self-reported measures. Against this backdrop, prior studies mainly employed self-reporting (e.g., survey data), and more direct measures of trust such as brain activity are only rarely found in the context of virtual human perception (e.g., the studies by Mustafa and Magnor (2016), Mustafa et al. (2017), and (Nissen et al., 2023) are of the few studies we could identify on this topic). The generalization of knowledge gained from studies on avatars in general (such as Riedl et al. (2014a)) to photo-realistic virtual influencers is limited because i) most avatar-focused research employs clearly computer-generated images and not avatars that aim at high realism, and ii) when looking at avatars, we usually look at a representation of an individual human being. In contrast, virtual influencers are not the representation of a human individual, but they are designed as their own fictional individual character. Therefore, mind attribution of virtual influencers, and resulting perceptions of trust, uncanniness, and social presence might differ significantly between avatars and virtual influencers. This is also related to the fact that iii) virtual influencers are active on a social media platform which is not limited to visual representations, but also include texts. We therefore raise the following research question:

RQ1: How do virtual compared to human influencers impact mind attribution and the perceptions (trust, social presence, uncanniness) of social media users?

In a second step, it stands to question if these evaluations differ depending on whether people really *realize* if the shown influencer is a human being or a virtual entity. In the broader social media literature, transparency is provided by disclosing sponsored ads or products (Djurica & Mendling, 2020). In previous research on virtual influencers, transparency has been implemented by disclosing that a social media influencer is not a real human, by providing a definition of a virtual influencer and by disclosing who created its content (Lim & Lee, 2023).

Results show that disclosure often leads to higher perceived trust, but there are also cases showing that this does not always have to be the case (e.g., warning messages for fake news) (Oeldorf-Hirsch et al., 2020; Ross et al., 2018). In practice, virtual influencers often self-disclose by describing themselves as ’robots’ or ’virtual beings’ which can cause confusion due to discrepancies between their human-like design and non-human identity (Cornelius et al., 2023). Since trust and transparency through disclosure are closely linked (Venkatesh et al., 2016), an understanding of the relationship between these concepts could contribute to a better understanding of virtual influencers. In addition, a lack of transparency could also have a negative impact on authenticity and thus, on the effectiveness of influencer marketing (Lim & Lee, 2023). Furthermore, from an ethical point of view, transparency is an important principle which needs to be considered to address questions such as responsibility of the virtual influencers’ contents (Robinson, 2020). Therefore, we pose the following second research question:

RQ2: How are mind attribution and the perception of virtual influencers affected by self-disclosure?

In order to answer both research questions, we conducted three studies. First, we conducted an online survey (study 1) with N = 112 participants focusing on the examination of the perceived trust, social presence, and uncanniness toward virtual and human influencers. From this study, we select two pairs of virtual and human influencers to investigate them in-depth in our second study (study 2). That is, we conducted a laboratory neuroIS experiment with N = 34 participants using functional near-infrared spectroscopy (fNIRS) in order to explore if there are also implicit neural processes in response to human and virtual influencers that may not necessarily translate into self-reported constructs. Through this, we identify neural mechanisms reflecting mind attribution processes that may add further explanation to how the phenomenon of virtual influencers is processed in decision making areas of the brain. As building and establishing trust can be challenging to implement and observe in virtual environment setting (Srivastava & Chandra, 2018), the identified neural mechanisms might be crucial antecedents to trust building. Third, we conducted an online survey in study 3 to validate our findings from study 1 and 2, by measuring self-reported mind attribution and by controlling for influencing factors such as ethnicity of influencers in comparison to the participants' ethnicity as well as perceived attractiveness and authenticity of the influencers.

We contribute to IS research by deriving knowledge on how people perceive highly human-like virtual influencers in social media in comparison to human influencers. We provide an explanation for the connection between trust, uncanniness, and social presence in the context of the ToM. We suggest how mind attribution and trust are connected in the context of a technological phenomenon in social media that gained increasing importance in IS and related disciplines. Furthermore, we provide initial guidance for developers of virtual influencers and for decision makers in organizations by revealing how businesses can successfully integrate virtual influencers in their social media communication strategies.

The remainder of this paper is structured as follows: First, we provide an overview of relevant previous work on trust and transparency in social media and introduce our theoretical background. Second, we derive our hypotheses and present the research design for both studies. Third, we summarize our results followed by a discussion of the findings. Lastly, we draw conclusions, show limitations and areas of interest for future research, and summarize our contribution to IS research.

2 Literature Background: Human-like Virtual Influencers

As one of the first successful virtual influencers in social media, Lil' Miquela appeared in 2016 and published her first post on Instagram. In 2022, she has nearly three million followers on Instagram and more than one million followers on other social media platforms such as TikTok and Facebook. Virtual influencers exist in nearly every country and culture. According to VirtualHumans.org, there were more than 200 virtual influencers on various social media platforms in 2022. They promote products of popular brands such as Prada, Porsche, Samsung, or Ikea (Hofeditz et al., 2023).

Artificial agents recommending products online are not a new phenomenon

(Benbasat & Wang, 2005; Kretzer & Maedche, 2018). However, virtual influencers as computer-generated, mostly fictional characters, with their "own" engaging social media accounts are a unique type of artificial agents with specific characteristics and increasing relevance for businesses (Lim & Lee, 2023). They are agents engaging in their own emotional social media content to gain and keep followers (Mirowska & Arsenyan, 2023) which gives them a wide reach (Moustakas et al., 2020). With recent technological advances, they achieve a high level of attractiveness (Mouritzen et al., 2023) and authenticity (H. Kim & Park, 2023). According to Choudhry et al. (2022), virtual influencers can be described as "computer-generated characters, many of whom are often visually indistinguishable from humans and interact with the world in a first-person perspective as social media influencers" (p.1). Virtual influencers are not only computer-generated entities, but also designed with human-like expressions and their own personalities which can be expressed in their pictures by wearing certain clothes, facial expressions or gestures (Naumann et al., 2009; Shevlin et al., 2003; Willis & Todorov, 2006). Even the style of the picture can influence how they are perceived as Qiu et al. (2015) suggested that selfies result in an attribution of agreeableness, conscientiousness, neuroticism, and openness. Virtual influencers can trigger social and emotional responses (Park et al., 2021) and adapt to common human social media usage behaviors, such as using emojis, posting emotional content, and pictures with real-life people and other virtual influencers or computer-generated entities. To shape their identity, they sometimes post an entire story from their "lives" (Arsenyan & Mirowska, 2021). The content of the virtual influencers is thereby a blend of human and computer inputs (Robinson, 2020). However, the companies which manage the virtual influencers' profiles rarely disclose details about the development (Choudhry et al., 2022; Robinson, 2020).

Virtual influencers can have an animal-like (e.g. Bee Influencer) (Choudhry et al., 2022), cartoon-like (e.g., Guggimon), thing-like (e.g. Nobody Sausage) or human-like (e.g., Lil Miquela) visual representation (Dabiran et al., 2022). They can represent an existing human as a computer-generated avatar or, as in most cases, represent fictional characters. In contrast to social bots, which can be described as automated social media accounts that simulate human communication and interaction on social media (Hofeditz et al., 2019), virtual influencers attempt to establish their own identity with visual representation (Moustakas et al., 2020).

In 2018, TIME Magazine listed the human-like virtual influencer Lil' Miquela as one of the 25 most influential people on the Internet (Times, 2018) and in 2023, about half of the most successful virtual influencers are designed as highly-realistic and human-like (Molenaar, 2022). Due to technological advancements in computer-generated imagery, some highly realistic human-like virtual influencers can cause confusion by social media users about how they and their posts are created (Cornelius et al., 2023), if the process is partly or fully automated, for example, by applying machine learning (Robinson, 2020), and even sometimes whether they are a real human or not (Hofeditz et al., 2023). Although the perception of these human-like virtual influencers by social media users raises relevant issues, it is still under-researched (Ozdemir et al., 2023). That is why we focus on photo-realistic human-like virtual influencers in this work.

3 Theoretical Background

3.1 Factors Influencing Trust in Computer-Generated Social Media Actors

Trust is a concept which has been addressed by a wide range of studies across a plethora of research disciplines and was already discussed by philosophers in the ancient world (Baier, 1986). In general, trust describes the relationship between a trustee and a trustor (Dunn, 2000). A look at the literature and its various discourses on trust and trustworthiness highlights the complexity of the concept, which follows different premises depending on the context and discipline in which it is discussed (Rousseau et al., 1998; Simpson, 2012). Trust in influencers involves confidence in the reliability and integrity of the perceived actor and the posted content on social media (D. Y. Kim & Kim, 2021). McKnight et al. (2011) distinguished between trust in people which can be defined as the interpersonal willingness to depend on another party because of certain characteristics, and trust in technology which focuses more on the functionality, perceived helpfulness, and reliability of a system. According to McKnight et al. (2011) trust in people is usually institution-based and more indirect than trust in technology, which relies more on a knowledge-base and functionality. Computer-generated avatars such as virtual influencers try to mimic human appearance and behavior (Lim & Lee, 2023) which might trigger predominantly but not exclusively human-like trusting beliefs.

Previous research found indications that trust evaluations are more difficult to make for avatars in comparison to human faces (Riedl et al., 2014a). This could cause uncertainty of how the agent can be evaluated correctly based on discrepancies between the perceived human-likeness and the non-human identity. Previous research suggests that this might result in less trust towards computer-generated influencers (Cornelius et al., 2023), which would mean higher trust in human influencers. We therefore hypothesize:

H1a: *Human influencers will be rated higher in perceived trust when compared to similar virtual influencers.*

Trust in social media content can be influenced by different factors such as perceived social presence, percieved personality (Naumann et al., 2009), needs or knowledge (Shareef et al., 2020). Social presence is defined as the perceived closeness and warmth toward a computer system, but also as the perceived degree to which a computer system is able to create warmth and closeness of other human users (Lu et al., 2016; Ogonowski et al., 2014; Short et al., 1976). Social presence is generated by factors such as face-to-face communication (Hess et al., 2009). Creating social presence in an online environment can be used to overcome the lack of warmth, social cues, and face-to-face interaction (Hess et al., 2009). With human-like computer-generated actors such as virtual influencers, these two perspectives of social presence begin to get more similar as this technology is a computer-generated system on the one hand, and acting on another computer system (a social media platform) on behalf of a human owner and controller on the other hand. Previous research already stated that virtual advisors and decision aids can increase social presence which is a key factor for creating trust in e-commerce (Pavlou et al., 2007). Since a human influencer usually has more and easier opportunities to generate social presence, for example by creating content in which they meet other humans in the physical world or by potentially being able to meet their fans, it can be assumed that human influencers are perceived as having a higher social presence than purely computer-generated virtual influencers. Although virtual influencers also sometimes present themselves in content showing them with people in the real world, this is usually connected with more effort and acting performance. We therefore assume the following:

H1b: *Human influencers will be rated higher in social presence when compared to similar virtual influencers.*

In contrast to social presence, which can increase trust in people or technology, uncanniness can reduce trust in a virtual actors (Mathur & Reichling, 2016; Nissen & Jahn, 2021). Uncanniness describes a person's feeling of something odd, mysterious or unexpected towards a system, a computer-generated avatar or a similar technology that gives the observer an uneasy feeling (Geller, 2008; Hanson et al., 2005; Mori, 1970; Mori et al., 2012). Mori et al. (2012) concluded that organizations should avoid creating too human-like avatars (Mori et al., 2012). Other scholars think that some computer-generated characters have already crossed the uncanny valley (Seymour et al., 2021), which would allow creating human-like avatars that cause even higher trust ratings than similar human avatars (Nightingale & Farid, 2022). In this work, we therefore focus on those virtual influencers that aim to look and behave highly human-like as it is unclear whether they cause perceived uncanniness or already have crossed the uncanny valley (Seymour et al., 2021). Although there are cases in which virtual influencers can be falsely classified as human, most virtual influencers currently do not look fully human-like and could cause uncanniness. A stronger feeling of perceived uncanniness towards virtual influencers in comparison to human influencers is therefore still likely. We therefore hypothesize the following:

H1c: *Human influencers will be rated lower in uncanniness when compared to* *similar virtual influencers.*

3.2 Mind Attribution Towards Computer-Generated Avatars

Perceived uncanniness is not limited to a feeling of eeriness towards the visual representation of a highly-realistic human-like artifical characters, but can also be felt if an artifical character behaves in a way that is interpreted as it would have its own intentions, emotions, and cognitive processes. In reference to the ToM, previous scholars described this as the uncanny valley of mind (Stein & Ohler, 2017).

ToM describes the ability of individuals to predict the behavior of others based on their knowledge, beliefs, and desires, i.e., their minds (Frith & Frith, 2010). This ability evolves during the first five years of childhood. Thus, children up to the age of five judge the knowledge, desires, and beliefs of others based on their own knowledge rather than on what another person could really know. From the age of five on, they then develop the ability to estimate one others' intentions, emotions, and cognitive skills; the process of which is called mind attribution (Kozak et al., 2006).

Previous research found indications for a two-systems account of mind attribution: a fast social-perceptual system, and a slower one based on reflexive cognitive operations (Meinhardt-Injac et al., 2018). Both systems can be considered as two classes of distinct processes (Keysers & Gazzola, 2007). The first social-perceptual system which is implicit and more automatic, immediate, and reflex-like, usually helps decoding socially relevant cues such as facial expressions, gaze direction, and body motion. The second system is more explicit and involves the cognitively demanding processes of mind attribution which is slower and reflective. It includes the explicit representation of mental states and beliefs (Meinhardt-Injac et al., 2018).

Mind attribution can be considered as the underlying process for other theoretical approaches in human-agent interaction in which people attribute human-like characteristics to non-human actors. Anthropomorphism as one of these concepts involves the tendency to attribute human-like characteristics (motivations, intentions, or emotions) to non-human agents (Epley et al., 2007; Waytz et al., 2010; Waytz et al., 2014). This can be triggered by certain anthropomorphic design features such as a human-like design (Cornelius & Leidner, 2021) but also certain behavior patterns (Stein & Ohler, 2017). Another theoretical approach is the computers-are-social-actors (CASA) paradigm which suggests that people show social responses to computers independently from their conscious beliefs (Nass et al., 1994). In summary, the CASA paradigm explores the way humans interact with technologies as if they were social actors, while anthropomorphism refers to the tendency to transfer human characteristics to non-human or technological entities (Bhatti & Robert, 2023; Gambino et al., 2020; Nowak & Fox, 2018). Both anthropomorphism and CASA are limited to the perception of and behavior towards non-human actors, and are based on the evolutionary mechanism of mind attribution.

To be able to make comparisons between the perception of human and virtual influencers, we therefore identified mind attribution as the fundamental underlying mechanism for other human-agent interaction explanation approaches such as anthropomorphism and CASA. We therefore use the term mind attribution for the process of ascribing a mind to both human and non-human actors (divided into perceived intentions, emotions, and cognition). We use the terms lower and higher mind attribution as referrals to the degree to which an agent is believed to have its own intentions, emotion, and cognition.

In line with our prior hypothesizing, the degree of mind attribution to an agent can have a significant impact on trust evaluations (Mou et al., 2020). Previous research found a positive correlation between selective trust and the ability of mind attribution (DiYanni et al., 2012) as we trust another actor based on our conception of their actions, beliefs, and intentions (Ruocco et al., 2021). The evaluations of an actor's intentions, emotion, and cognition, and their relevance for the interaction are processed on a neural level in the human brain.

One brain area to which such processes can be attributed to is the medial PFC (mPFC) (Lieberman, 2007; Molenberghs & Morrison, 2014). The mPFC has evolved to evaluate intentions and states of mind of others, and to calculate their relevance for one's own well-being (Frith & Frith, 2010; Lieberman, 2007; Satpute et al., 2014; Weaverdyck et al., 2021). Several works suggested a possible relation between higher mind attribution and activation in the mPFC (Lieberman, 2007; Molenberghs & Morrison, 2014). According to ToM, the evaluation of the states of others (i.e., mind attribution) is based on a hereditary predisposition to recognize human features (for e.g., faces, body movement) (Saxe & Baron-Cohen, 2006). As a result of this, neural processes which might be related to mind attribution are thought to be related to the level of human-likeness in artificial agents such as virtual influencers. This reasoning is supported by neuroscientific findings stating that human-like agents lead to increased activation of the mPFC compared to less human-like agents (Krach et al., 2008; Miura et al., 2009). As a result, researchers have claimed that the mPFC may act as an encoder for human-likeness, and respective perceived uncanniness (Rosenthal-Von der Pütten et al., 2019; Wang & Quadflieg, 2014). This is also supported in the mPFC's role in encoding both approach-intentions to positive events, as well as avoid-intentions to highly uncertain events (Burgos-Robles et al., 2017; Etkin et al., 2011; Xue et al., 2009).

More precisely, literature supposes that positive and negative events are processed on different hemispheres. This so called valence lateralization hypothesis proposes that positive events are processed on the left hemisphere, while negative events are processed on the right hemisphere (Davidson, 1992, 1998; Sackeim et al., 1978; Wager et al., 2003). Across different contexts, several works found support for this hypothesis with regard to mPFC activation (Killgore & Yurgelun-Todd, 2007; Li et al., 2019; Nissen & Krampe, 2021; Wager et al., 2003).

Summarizing the presented knowledge, we assume that humans are better able to judge and attribute mind to real humans compared to artificial agents, including virtual influencers. Therefore, we hypothesize that the mind attribution of human influencers will be higher than for virtual influencers:

H2a: *The perception of virtual influencers will lead to less mind attribution compared to the perception of similar human influencers.*

While mind attribution is lower for virtual influencers, we also proposed that they will be associated with higher uncanniness. Therefore, difficulties that human have in attributing a mind to the virtual influencers may result in a negative affect toward these influencers, which may also become evident in their higher perceived uncanniness. Given that especially the right mPFC area involves the evaluations of mind attribution (Davidson, 1992, 1998; Wager et al., 2003), as well as the encoding of the negative valence of such evaluations, we hypothesize that the right mPFC will be activated for virtual compared to human influencers:

H2b: *The perception of virtual influencers will lead to higher right mPFC activation compared to the perception of similar human influencers.*

3.3 Possible Effects of Self-Disclosure of Virtual Influencers

Social media users can get confused by the level of realism of human-like virtual influencers and sometimes even fail to correctly distinguish between virtual and human influencers (Batista & Chimenti, 2021). Robinson (2020) suggested to discuss whether virtual influencers should be labeled as such to address ethical issues related to unclear moral responsibility and low transparency. Some virtual influencers already disclose themselves as robot or being virtual (Ahn et al., 2022; Cornelius et al., 2023). Previous research proposed an impact of disclosing advertisements in social media on perceived trustworthiness of influencers and purchase intention (Djurica & Mendling, 2020). Djurica and Mendling (2020), however, stated that research on the effect of disclosure on trust ratings in the context of social media influencers is missing. We therefore think that examining the effect of disclosure is important. We assume that it is likely that virtual influencers that disclose themselves as non-human are rated higher in trust in comparison to those that do not disclose themselves as virtual influencers. We hypothesize:

H3a: *Undisclosed virtual influencers will lead to lower perceived trust ratings than disclosed virtual influencers.*

Self-disclosure is usually implemented by adding textual information in the influencer's profile or in a single post's caption such as referring to themselves as robots (Cornelius et al., 2023). Previous research suggested that socially rich textual information of virtual recommendation agents can increase the perceived social presence (Hess et al., 2009). In addition, research on conversational agents could show how social cues such as textual self-disclosure can positively affect social presence (Feine et al., 2019). We therefore assume the following:

H3b: *Undisclosed virtual influencers will lead to lower perceived social presence ratings than disclosed virtual influencers.*

Regarding the uncanny valley effect, previous research has not shown a consensus on the effects of transparency. Stein and Ohler (2017) suggested that disclosing an avatar as having a mind on its own and behaving highly human-like can increase perceived uncanniness. In contrast, Skjuve et al. (2019) assume that a lack of self-disclosure leads to more perceived uncanniness in chatbots. In context of virtual influencers, it stands to question how the disclosure of the influencer's nature impacts user perceptions as there is no available research yet. In the broader frame of advertisement disclosure on social media, disclosing product placements has shown to have a negative impact on the attitude toward the post and the influencer (Boerman et al., 2017; Karagür et al., 2022). These findings, together with the claims of Stein and Ohler (2017) give the indication that when human-like virtual influencers become almost indistinguishable from humans, the actual reveal of them not being human might lead to a more negative affect in the sense of higher perceived uncanniness. Reason for this may be that while having conflicts in mind attribution when not disclosed, we still try to attribute a mind (Meinhardt-Injac et al., 2018). Upon disclosure it becomes clear that the presented influencer is computer-generated, and does not have a mind on its own. As a result, it disguises as something that it is not, resulting in a feeling that something is off, and therefore, higher uncanniness. We therefore assume the following hypothesis for virtual influencers:

H3c: *Undisclosed virtual influencers will lead to lower perceived uncanniness ratings than disclosed virtual influencers.*

Mind attribution is assumed to be made based on implicit and explicit processes (Frith & Frith, 2010). While explicit processes involve the cognitive-demanding evaluation of the representation of others' mental states and beliefs, implicit processes include the faster and more automated transmission of information by signals such as facial expressions, gaze direction, vocalization or body motion (Meinhardt-Injac et al., 2018).

Since virtual influencers have such signals as facial expressions, gaze direction and often vocalization (Dabiran et al., 2022), it is likely that implicit processes are triggered in the spectator. If it is not disclosed that it is a virtual influencer, this may lead to uncertainty in mind attribution (Meinhardt-Injac et al., 2018). This is decided in favor of the implicit processes, especially when influencers are viewed just briefly, as it is common in social media. This might result in higher mind attribution in comparison to disclosed virtual influencers. However, if it is disclosed that it is a virtual influencer, uncertainty might be reduced which results in the viewer correctly judging mind attribution by comparing implicit and explicit processes (Meinhardt-Injac et al., 2018). This could lead to a lower mind attribution for disclosed virtual influencers, since it can be judged more explicitly that it is not a human being. We hypothesize:

H4a: *Undisclosed virtual influencers will lead to higher mind attribution compared to disclosed virtual influencers.*

As self-disclosure of virtual influencers might reduce uncertainty, but also reveal them more clearly as non-human, we assume that mPFC activation is higher for undisclosed virtual influencers. In line with our prior theorizing for H4a, it is likely that a higher mind attribution takes place when participants see the virtual influencer as a highly human-like agent without without any further description. Due to the higher mind attribution and lower uncanniness of the undisclosed influencers, it is likely that the neural processing of the signifies positive valence. Therefore, we believe in accordance with the prior introduced valence lateralization hypothesis, that the left mPFC will be activated for undisclosed over disclosed virtual influencers (Davidson, 1992, 1998; Sackeim et al., 1978; Wager et al., 2003). We therefore hypothesize that:

H4b: *Undisclosed virtual influencers will lead to higher left mPFC activation compared to disclosed virtual influencers.*

4 Method

In order to better understand differences in the perceptions of virtual and human influencers and the effect of self-disclosure on mind attribution, trust, social presence, and uncanniness, we conducted an online survey study with N = 112 participants that we used as a study 1, and a laboratory experiment with N = 34 participants using fNIRS as a direct measurement approach as study 2. Using both approaches allowed us to not only observe explicit self-reports, but also directly measure implicit neural processes that can be attributed to mechanisms relevant for the formation of perceived trust. Finding support for our hypotheses in study 2, we validate this support in an online experiment that controls for additional influencing factors such as authenticity, attractiveness, and ethnicity, while also assessing mind attribution as self-reported measure (study 3).

5 Study 1: Behavioral Experiment for Stimuli Selection

5.1 Sample

A sample of $N = 112$ participants was recruited through clickworker in German speaking countries. First, participants provided demographic data, as well as information about the usual Instagram usage at first. 43.8% of the participants are female (all remaining are male), and the age ranged from 20 to 66 years ($M = 35.5, SD = 11.1$). Regarding their social media activity, 3.6% stated that they do not use Instagram, 11.6% assume they use it several times per month or less, 15.2% use it at least once a week, and 69.6% state that they use Instagram at least once a day, with the majority using it several times per day. Regarding their interest in influencers, 29.5% of the sample stated that they do not follow any influencers on Instagram, 64.3% stated that they follow some (semi-professional) influencers, and only 6.3% stated that they would follow many influencers on Instagram.

5.2 Stimuli

In order to identify suitable stimuli for this study, an extensive search through operating virtual influencers on Instagram has been conducted by four independent coders. The target were virtual influencers that had at least 20,000 followers and aimed to represent a highly human-like (photo-realistic) influencer whose Instagram profile mimics that of comparable human influencers. We further targeted a diverse set of influencers with regard to their ethnicity. Therefore, we selected one virtual representative of four major ethnic groups: i) Asian, ii) (Northern) European, iii) Latin (American), and iv)

Account name	Ethnical group	Influencer type	Follower count on March 10th 2022
bermudaisbae	(Northern) European	Virtual influencer	>276 thousand
dagibee	(Northern) European	Human influencer	>6.5 million
shudu.gram	African	Virtual influencer	>226 thousand
anokyai	African	Human influencer	>709 thousand
zoedvir	Latin (American)	Virtual influencer	>27 thousand
marinadnery	Latin (American)	Human influencer	>46 thousand
rozy.gram	Asian	Virtual influencer	>123 thousand
jenndeugikim	Asian	Human influencer	>405 thousand

Table 1: Included Instagram Accounts for Stimuli Selection

African. For each of the virtual influencers, we then searched for similar looking human influencers on Instagram. This procedure was repeated until all four coders reached consensus on the selected influencers. The resulting set of influencers is listed in Table 1.

To avoid potential bias due to different color tones, all of the selected Instagram posts were shown in greyscale. The selected influencer posts are depicted in Figure 1.

5.3 Measures and Study Design

The procedure of the questionnaire is as follows. First participants were welcomed to the study and briefed about the study's purpose. After that came demographic questions, as well as questions related to Instagram use behavior. This is followed by a randomly selected order in which each influencer post was shown together with questions related to the perceived trust (adapted from D. Y. Kim and Kim (2021)), perceived social presence (adapted from Gefen and Straub (1997)), and perceived uncanniness of the influencer (adapted from Tinwell and Sloan (2014)). Additionally, we asked for the perceived humanness (adapted from Holtgraves and Han (2007)) as manipulation check for the degree to which participants were able to tell human and virtual influencers

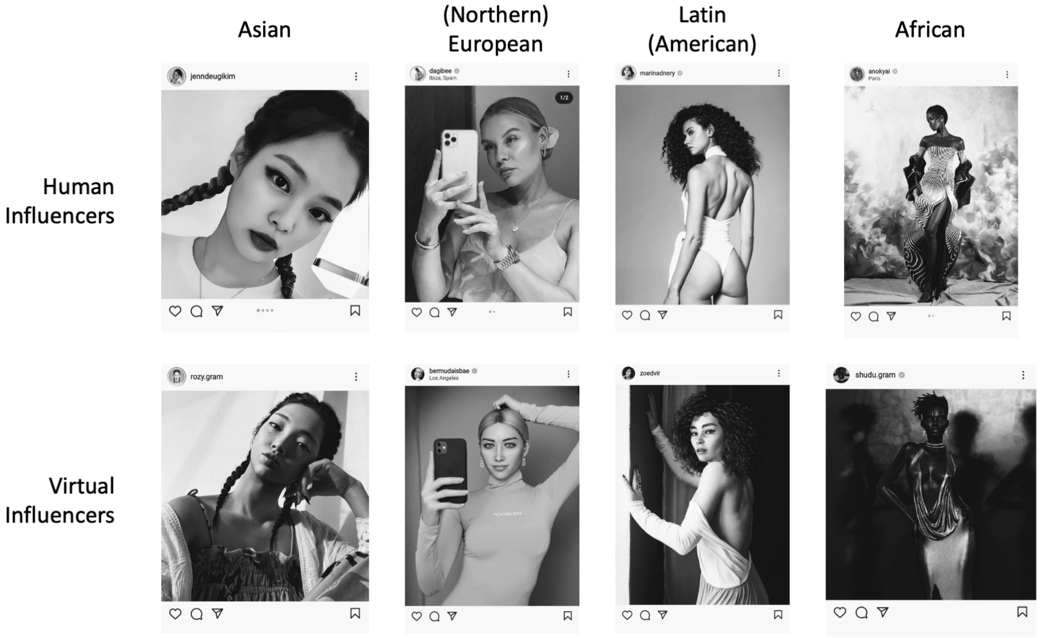


Figure 1: Selected Stimuli

apart.

5.4 Results

We included a manipulation check for each influencer in which participants had to evaluate whether the shown influencer was virtual or human. In the following Table 2, the part of the sample who believed that the shown influencer is human is provided. We can see here that at least 50% of the sample thought of the presented influencer post to be human, even if it was virtual. Consequently, at least half of our participants were not able to distinguish between virtual and human influencer. The difference between the virtual and human Asian influencer was most diminishing as the same amount of the sample thought of both as being human. The Friedmann test show that although at least half of the sample mistake the virtual influencers as humans, the number of

	Virtual*	**Human***	χ^2**	**p**
Latin	63.40%	80.40%	11.8	<.001
African	50%	74.10%	22.3	<.001
European	65.20%	92%	28.1	<.001
Asian	65.20%	65.20%	0.026	.873
**% of sample that believes shown Influencer is a human **Friedmann Test*				

Table 2: Manipulation Check ”Is this influencer human?”

right answers for the human influencers was significantly higher (except for the Asian influencers).

Nevertheless, to test whether participants did evaluate the human and virtual influencers differently regarding the hypothesized constructs of trust, social presence, and humanness, repeated-measures one-way ANOVAs were calculated. Results reveal that although a great part of participants could not correctly distinguish between virtual and human influencer, the virtual influencers were rated significantly lower in trust ($F(1, 111) = 5.63, p = .019, \eta_p^2 = .048$). Further, the virtual influencers were also rated significantly lower in their social presence ($F(1, 111) = 4.84, p = .03, \eta_p^2 = .042$), and their humanness ($F(1, 111) = 9.181, p = .003, \eta_p^2 = .076$) compared to the included human influencers.

6 Study 2: In-Depth Investigation of Selected Influencers with fNIRS

6.1 Sample

For study 2, N = 34 participants were recruited from the local University that used Instagram at least once a month (2.9%). Most participants used the platform several times a day (76.5%), only few used Instagram more frequently (i.e., several times per hour, 8.8%), and a similar number of participants used it less frequently (i.e., at least once a week 11.7%). Average age was $M = 24.5$ years ($SD = 4.2, Min = 19, Max = 34$). About half of the sample were male (41.2%, 58.8% were female), and the majority of the sample qualifies as right-handed (91.2%) as measured with the German version of the laterality quotient by Salmaso and Longoni (1985).[1] Regarding their employment, the majority of recruited participants were students (82.4%), with the remaining participants being employed (17.6%).

Regarding their interaction with social media influencers, 5.9% say that they follow a lot of influencers, 79.4% state that they follow some influencers, and 14.7% claim that they do not follow any influencers. Because prior research has shown that the dispositions to trust can impact trust itself, we have included personal trust in technology (PTT), and dispositional trust to influencers (DTI) as control variables. Results show that on a 7-point Likert-scale, PTT is closer to the higher end of the scale ($M = 4.78, SD = 1.15$), while DTI is on the lower end of the scale ($M = 2.29, SD = 0.796$). Consequently, the recruited sample is highly confident in using technology, but seems to have lower trust in social media influencers in general.

6.2 Stimuli

As stimuli, comparable posts on Instagram from 4 different influencers from study 1 were selected, two of which are human and two are virtual influencers (i.e., DagiBee (human), bermudaisbea (virtual), jenndeugikim (human), rozygram (virtual)). The

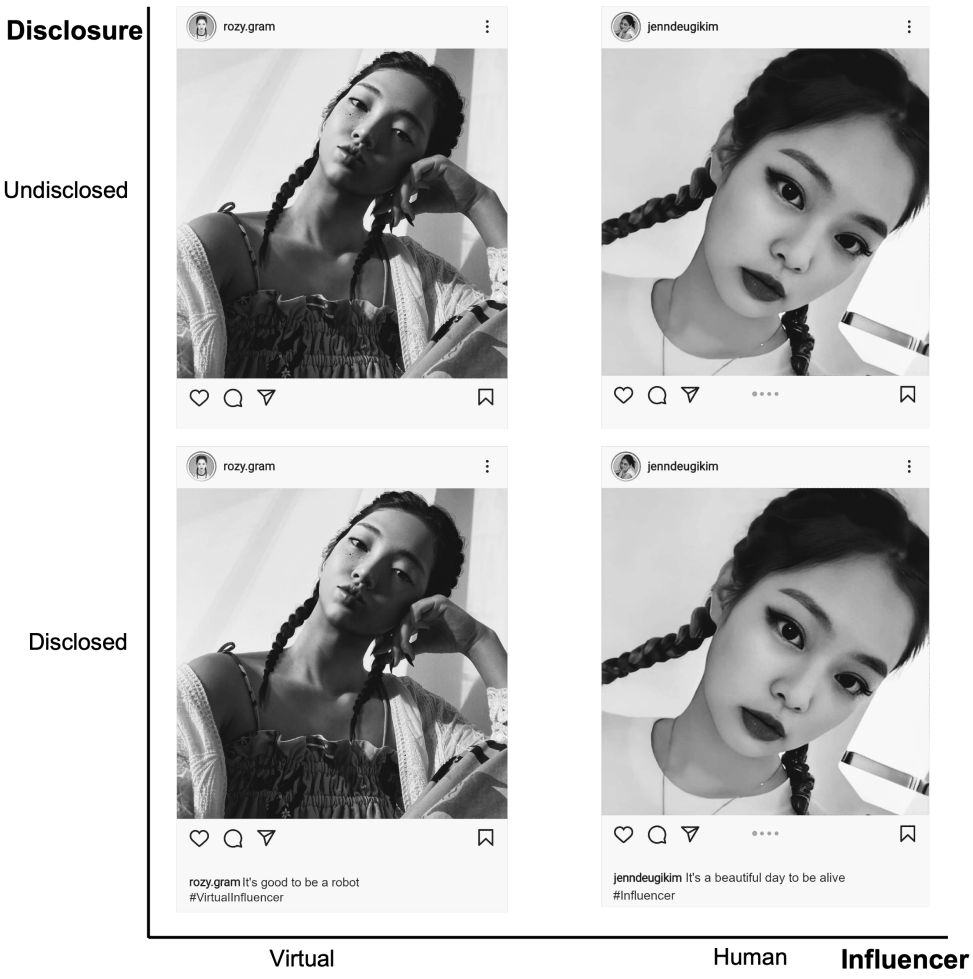
[1] Note that handedness has long been thought to impact the hemispheric lateralization of neural activity in the PFC, which often lead to exclusion of left-handed people as participants. However, including left-handed people accounts for the natural diversity in humans (Cinciute et al., 2018; Schmitz et al., 2017). Therefore, we did not recruit participants based on their handedness, but only used it as another demographic characterization of the sample.

influencers and posts have been selected from the results of study 1 based on the reported perceived humanness. That is, we selected the human - virtual influencer pair that had the highest difference in perceptions of humanness (i.e., DagiBee and bermudaisbae) and the human-virtual influencer pair that had the lowest difference in perceived humanness (i.e., jenndeugikim and rozygram). Further, related works with fNIRS measurements have shown that changing the color of the stimulus leads to differences in neural processing of the stimulus (Nissen, 2020), which in our case, is not targeted. Therefore, to avoid possible bias due to different coloring of background and clothes, all posts were shown in gray-scale.

This time, not only the comparison between human and virtual is of interest, but also whether the perception differs depending on the disclosure of the influencer. Therefore, for each included influencer post, we have two versions: one of which does not include additional information about the influencer, and one of which discloses the influencer as virtual or human. An example for one human-virtual comparison is shown in the following Figure 2.

6.3 Measures and Study Design

The overall procedure of the study was as follows: first, the participant was welcomed at the lab and placed in a seat in front of a desk with the experiment display and keyboard. After that, the participant was handed detailed study instructions and a consent form; both of which s/he was asked to carefully read. In case a participant had any remaining questions, they were answered by the experimenter. With the participant having given informed consent, the experimenter opened a questionnaire including demographic questions, handedness, as well as personal trust in technology scales (taken from



rozy.gram It's good to be a robot
#VirtualInfluencer

jenndeugikim It's a beautiful day to be alive
#Influencer

Figure 2: Example Stimuli 2 (virtual vs. human) x 2 (undisclosed vs. disclosed)

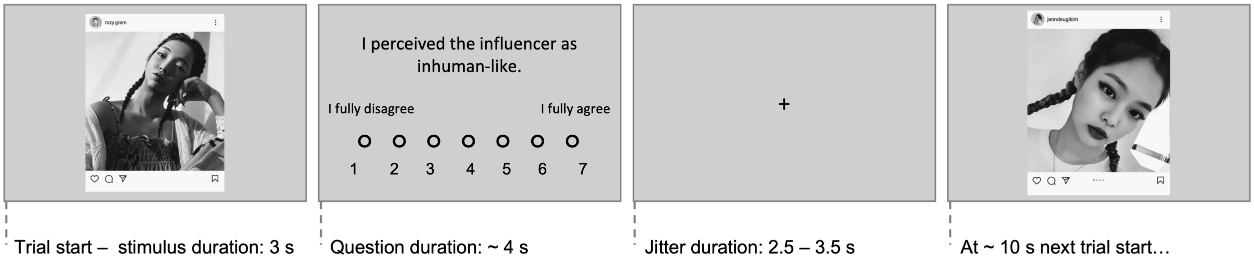

I perceived the influencer as inhuman-like.
I fully disagree
I fully agree
○ ○ ○ ○ ○ ○ ○
1 2 3 4 5 6 7
+
Trial start – stimulus duration: 3 s
Question duration: ~ 4 s
Jitter duration: 2.5 – 3.5 s
At ~ 10 s next trial start…

Figure 3: Experimental Paradigm

McKnight et al., 2011; McKnight et al., 2002) and dispositional trust in influencers scales (D. Y. Kim & Kim, 2021) .

After that, the fNIRS headband was placed on the participant's head and calibrated. This is followed by the experimental paradigm in which selected influencers are shown. More precisely, we used a 2x2 factorial, within-subjects design with influencer (human, virtual) and disclosure (undisclosed, disclosed) as independent variables. As more fully discussed in the related literature, both the nature of the influencer and whether users are given full information about it may heavily impact their perception of the influencer in terms of trust, social presence, and uncanniness. For this reason we included self-reported scales of two items for trust adapted from D. Y. Kim and Kim (2021) and Kennedy et al. (2001), five items for social presence adapted from Gefen and Straub (1997), and four items for uncanniness adapted from MacDorman et al. (2009) and Tinwell and Sloan (2014).

The described stimuli and scales were included in an event-related experimental design. In the first half of the design, participants saw one of the undisclosed influencer posts for 3s, after which one of the questions from the dependent variables appeared together with a 7-point Likert scale reaching from 1 = "I fully disagree" to 7 = "I fully agree". An eye tracking study has shown that the fixation duration on Instagram posts lies between 2 - 5s, which is why the 3s duration for stimulus presentation was selected (Zhou & Xue, 2021). The order in which the influencers and questions were shown was completely randomized.

After the input of the rating, a cross jittered between 2.5 - 3.5s was shown as neutralizing image before the next trial began. After each of the questions were asked for each of the undisclosed influencer posts, a longer jittered cross appeared after which the whole procedure repeated for the disclosed posts.

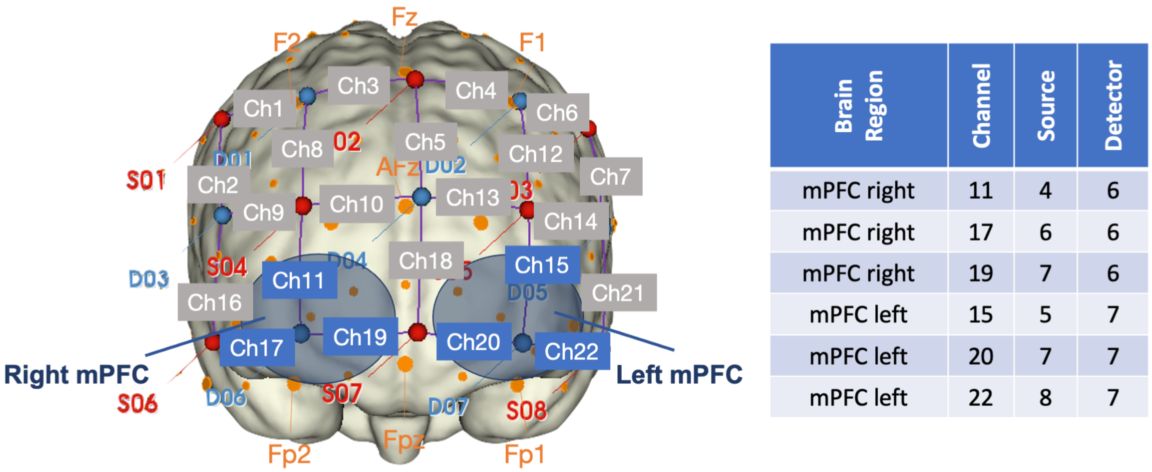
It has to be noted that this number of repetitions is necessary in brain imaging studies as a means to ensure that the elicited brain activation is a result of stimulation and not a false positive. We also need to bear in mind that the neurophysiological responses underlie natural variance between individuals which makes between-subjects designs difficult to design when avoidance of false positives is to be ensured. Therefore, while knowing of possible biasing effects due to the number of repetitions of stimulus presentation, it is still the validated and rigorous experimental setup for neuroimaging studies in the broader neuroscientific literature (Luck, 2014). To adhere to these standards, we selected this procedure for our fNIRS study as well.

As a result of this within-subjects design, each participant first saw all included influencers without disclosure, and in the second part saw all influencers with disclosing text. After both parts were finished, the fNIRS headband was removed from the participant's forehead and they were thanked for their participation. The study procedure was agreed upon by the local University's ethical committee.

6.4 Data Acquisition

We used a mobile, continuous wave NIRSport 1 device developed by NIRX for the acquisition of brain data. Technically, fNIRS sends near-infrared light into the skull at two (or more) wavelengths that are reflected or absorbed by the oxygen Mets of hemoglobin in the blood (Ferrari & Quaresima, 2012; Pinti et al., 2020). Specifically, we assessed the levels of the oxygenated and deoxygenated hemoglobin (HbO and HbR, respectively) in the brain regions under the fNIRS device (Krampe et al., 2017; Pinti et al., 2020). Both the HbO and HbR signals provide information identifying the brain region that is required to process the presented images of virtual influencers, because an increase in HbO levels and a decrease in HbR reveals that more oxygen is required in the respective brain area that signifies a neural activation (Pinti et al., 2020). Both the HbO and HbR signals are highly correlated to the BOLD signal produced by fMRI (Hoshi et al., 2001; Huppert et al., 2006; Noah et al., 2015; Pinti et al., 2020; Strangman et al., 2002; Toronov et al., 2003; Wijeakumar et al., 2017), which makes the results obtained by fNIRS strongly comparable to BOLD results from related work that used fMRI as a measuring method.

The utilized device in this study comes with a sampling frequency of 7.81Hz and has two wavelengths with 760nm and 850nm. For this study, measurements were focused on the PFC, which was covered with 8 sources, 7 long-distance detectors (LDD, average distance set to 30mm), and 8 short-distance detectors (SDD, average distance set to 8mm, one short distance detector for each source). The SDD are used to assess task-unrelated, extracerebral activation in the fNIRS signal. Thus, they provide an accurate measure to filter out noise in the data and help to ensure that only task-dependent



Brain Region	Channel	Source	Detector
mPFC right	11	4	6
mPFC right	17	6	6
mPFC right	19	7	6
mPFC left	15	5	7
mPFC left	20	7	7
mPFC left	22	8	7

Figure 4: Utilized fNIRS Channels for the Left and Right mPFC

neural activation is considered in the data analyses (Brigadoi et al., 2014; Goodwin et al., 2014; Yücel et al., 2016). Overall, the fNIRS montage holds 22 channels which cover most cerebral areas of the PFC. As we are interested in activation of the mPFC, we focus on channels which relate to medial areas of the PFC only.

The utilized montage of the fNIRS device is depicted in Figure 4. It identifies the LDD channels that cover the specific brain regions, with the channels that are used for data analysis colored blue. For better comprehensibility, the sources, detectors, and channels that make up the mPFC areas are also listed in the table (i.e., channels 11, 17, and 19 for the right mPFC, and channels 15, 20, and 22 for the left mPFC).

6.5 Data Analysis

We analyzed raw fNIRS data in Matlab by using the Brain AnalyzIR toolbox developed by Santosa et al. (2018). As a first pre-processing step, we removed all over- and non-saturated channels from the data. This was followed by a resampling of the sampling frequency to 4Hz, which helps to address the high autocorrelation in the fNIRS signal (Huppert, 2016). After that, we calculated optical density, followed by cleansing data with the included SDD channels using Linear Minimum Mean Square Estimations to filter out artefacts due to respiration, heart rate, Mayer waves, movements, and extracerebral activity (Saager & Berger, 2005; Scholkmann et al., 2014). Finally, we calculated hemoglobin values (HbO, HbR) by using the modified Beer-Lambert Law with a partial pathlength factor set to .10 (Delpy et al., 1988; Kocsis et al., 2006). For the hemoglobin values, we initially calculated hemoglobin changes per channel for each condition on a subject level. For this calculation, a general linear model (GLM) with the hemodynamic response function (hrf) as baseline and the autoregressive, iteratively reweighted least-squares (AR-IRLS) algorithm is used. We selected the AR-IRLS algorithm because it has shown to provide an accurate means to further correct for motion artefacts and serially correlated errors in the hemoglobin values in fNIRS measurements (Barker et al., 2013; Huppert, 2016). After that, we applied a mixed-effects model for the group analysis that uses a covariance weighted regression based on the results of the prior described subject-level GLM. In this model, the influencer posts are used as fixed effects, and individual differences of participants are treated as random effects. The following section presents results from the mixed-effects model that survived the threshold of $p < .05$, and false discovery rate corrected p-values p_{FDR} <= $.1$ (Benjamini & Hochberg, 1995). Furthermore, HbO and HbR of the same brain region have to show opposing effects, as an increase in HbO is always accompanied by a decrease in HbR, and vice versa. In case where both HbO and HbR point to the same direction, it is likely a false positive and should not be treated as actual effect in the data (Scholkmann et al., 2014).

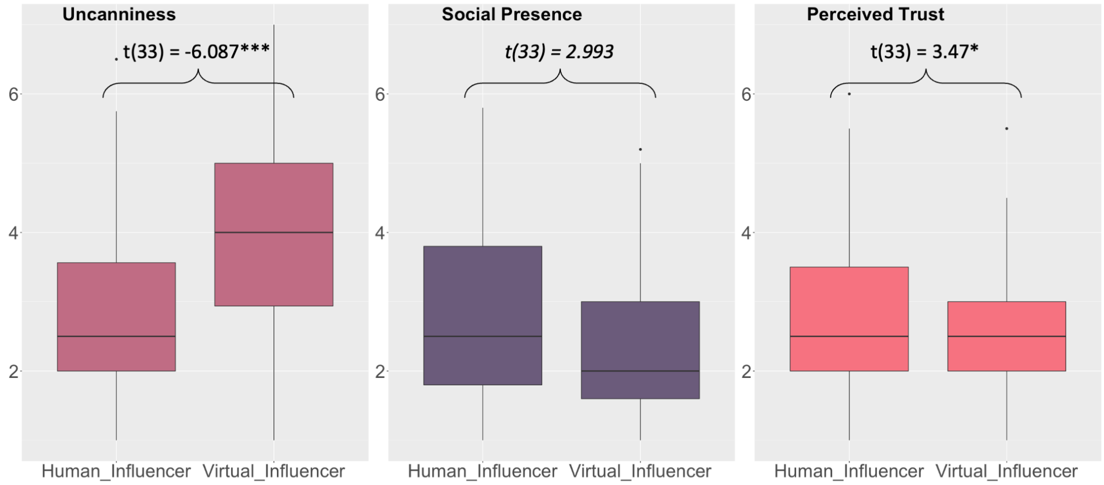


Figure 5: Self-Reported Results of Virtual vs. Human Influencers
Note. $*p_{tukey} < .05$, $**p_{tukey} < .01$, $***p_{tukey} < .001$

6.6 Results

6.6.1 Differences in Virtual vs. Human Influencer Perception

Self-Reported Results. For the included self-reported results in the experimental paradigm, a repeated-measures ANOVA was calculated. Overall, a significant effect could be identified for the human - virtual comparison across all included constructs ($F(1, 33) = 3.28, p = .079, \eta_p^2 = .09$), where trust and social presence are rated higher for the human influencers, while uncanniness is rated higher for the virtual influencers. In the post-hoc tests, it is revealed that the difference in uncanniness ($t(33) = -6.087, p_{tukey} < .001$) is by far the greatest, and differences in trust perceptions ($t(33) = 3.47, p_{tukey} = .017$) are greater than differences in social presence ($t(33) = 2.993, p_{tukey} = .054$). Thereby, hypotheses H1a - c are supported by these results. The results are visualized in Figure 5.

ROI	**type**	**Contrast**	$Beta$	SE	DF	T	p	p_{FDR}	$power$
right mPFC	hbo	Human - Virtual	-1.508	0.894	297	-1.687	0.093	0.111	0.315
right mPFC**	hbr	Human - Virtual	1.143	0.377	297	3.034	0.003	0.006	0.706
left mPFC	hbo	Human - Virtual	-1.492	1.218	297	-1.225	0.222	0.665	0.350
left mPFC	hbr	Human - Virtual	0.401	0.606	297	0.662	0.509	0.800	0.437

Table 3: fNIRS Results for the Comparison Between Human and Virtual Influencer
Note. $*p_{FDR} < .05$, $**p_{FDR} < .01$, $***p_{FDR} < .001$

Neural Results. In the comparison of neural activation in the medial PFC areas, only the right mPFC showed a significant higher neural activation for the virtual compared to the human influencers in the HbR signal ($beta = 1.143, t(297) = 3.3034, p_{FDR} < .006$). Notably, although not significant, we have the opposite effect in the HbO signal, supporting that this is actually an observed effect ($beta = -1.508, t(297) = -1.687, p_{FDR} = .111$).[2] Thereby, hypothesis H2 is supported.

6.6.2 Perception and Processing of Disclosure in Influencers

Self-Reported Results. Over all of the included influencers, no significant effect of disclosure could be identified across the included variables of trust, social presence, and uncanniness for disclosed over undisclosed influencer posts ($F(1, 33) = 2.639, p = .121, \eta_p^2 = .071$). When running the post-hoc tests for virtual influencers, we did only find a significant effect for uncanniness ($t(33) = -4.011, p_{tukey} = .014$), but not for trust ($t(33) = 2.507, p_{tukey} = .370$), nor for social presence ($t(33) = 1.167, p_{tukey} = .988$). Thereby, hypothesis H3a and H3b need to be rejected, while H3c is supported. The results are depicted as barplots in Figure 6.

Neural Results. In the comparison between the undisclosed and the disclosed

[2] HbO and HbR of the same brain region need to show opposing effects. An increase in HbO needs to be accompanied by a decrease in HbR (and vice versa) to enable us talking about an actual effect and not a false positive. In case where both HbO and HbR point to the same direction, it is likely a false positive and should not be treated as actual effect in the data (Scholkmann et al., 2014).

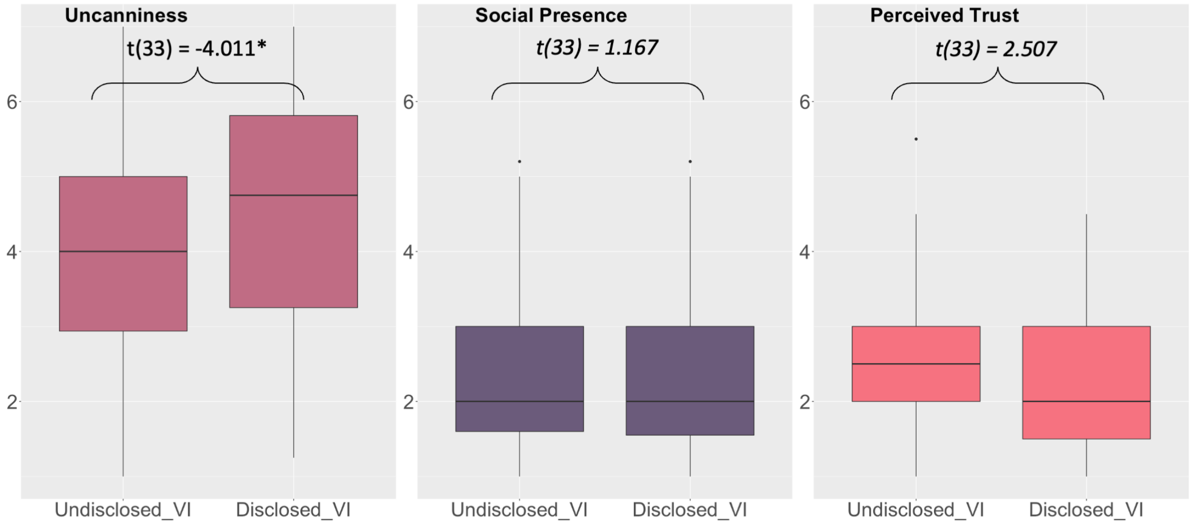


Figure 6: Self-Reported Results of Disclosed VS. Undisclosed Virtual Influencers
Note. $*p_{tukey} < .05$, $**p_{tukey} < .01$, $***p_{tukey} < .001$

virtual influencers, it does seem like there is a significant neural activation for the undisclosed influencers in the right mPFC in both the HbO and HbR signal (HbO: $beta = -1.725, t(297) = -2.675, p_{FDR} = .014$, HbR: $beta = -0.749, t(297) = -2.672, p_{FDR} = .014$). However, it has to be noted that both the HbO and HbR signal point into the same direction and that this is likely a false positive (Scholkmann et al., 2014). Therefore, we will *not* consider this activation an effect. Although less significant, there is an effect in the left mPFC HbR signal ($beta = -1.086, t(297) = -2.47, p_{FDR} = .084$), which is also supported in an opposing, non-significant HbO trend ($beta = 0.068, t(297) = 0.078, p_{FDR} = .938$). Therefore our hypothesis H4b is supported.

ROI	**type**	**Contrast**	$Beta$	SE	DF	T	p	p_{FDR}	$power$
right mPFC	hbo	Virtual(Undisclosed - Disclosed)	-1.725	0.645	297	-2.675	0.008	0.014	0.572
right mPFC	hbr	Virtual(Undisclosed - Disclosed)	-0.749	0.280	297	-2.672	0.008	0.014	0.571
left mPFC	hbo	Virtual(Undisclosed - Disclosed)	0.068	0.876	297	0.078	0.938	0.938	0.781
left mPFC*	hbr	Virtual(Undisclosed - Disclosed)	-1.086	0.440	297	-2.470	0.014	0.084	0.491

Table 4: fNIRS Results for Undisclosed (= NoCaption) vs. Disclosed (= Caption) Influencers
Note. $*p_{FDR} < .10$, $**p_{FDR} < .01$, $***p_{FDR} < .001$

7 Study 3: Online Experiment for Result Validation

7.1 Sample

To further test our hypotheses and validate the results suggested by the neural results in study 2, we conducted a third online survey via prolific. Overall, we distributed the questionnaire to 200 participants. 7 participants were filtered based on wrong answers to our control question, resulting in a sample of N = 193 participants. Average age of participants was $M = 30.6 years (SD = 9.11, Min = 18, Max = 65)$. Gender was equally balanced between male (49.2%) and female (48.7%) participants, with the remaining participants identifying as non-binary or preferring not to state their gender (4%). About half of the sample held a bachelor's degree (47.2%), followed by 23.3% having a postgraduate degree, 21.8% holding a high school degree or GED, 6.7% having a college or post-secondary certificate, and 1% holding no educational degree.

While we recruited mostly German participants in study 1 and 2, we were now interested to see whether our hypotheses hold true for an ethnically diverse sample. Major ethnic groups in the sample included White people (33.7%), Asian (24.4%), Black or African American (23.3%), Hispanic, Latino, or Spanish (16.1%). The remaining participants identified themselves with some other ethnicity (2.5%).

Regarding their social media activity, 11.9% stated that they use Instagram less than on a monthly basis, 6.7% assume they use it several times per month, 19.2% use it at least once a week, and 52.8% state that they use Instagram at least once a day, and 9.3% use Instagram on an hourly basis. Regarding their interest in influencers, 24.9% of participants follow less than 5 influencers, 15.5% follow more than 5, 22.3% follow more than 10, 18.7% follow more than 50, and 18.7% follow more than 100 influencer accounts. When asking for familiarity with the included influencers specifically, the majority of the sample (88.1%) did not know any of the included influencers. Each influencer was known by less than 5% of participants, ensuring that we do not have biased data due to high familiarity with some of the influencers (dagibee: 3.6%, bermudaisbae: 3.1%, shudu.gram: 3.6%, anokyai: 2.1%, zoedvir: 2.6%, marinadnery: 0.5%, rozy.gram: 2.1%, jenndeugikim: 0.5%).

7.2 Stimuli, Measures, and Study Design

Stimuli. The included stimuli are all 8 influencers presented in Figure 1 for Study 1, both with and without the captions exemplified in Figure 2 for Study 2.

Study Design. The questionnaire started with a welcome page on which participants were informed about the topic of the study and the type of questions that are to be asked. Upon providing their consent to participate, participants were asked demographic questions, as well as questions about their Instagram use behavior reported above in the sample section. After that, the influencers were presented in randomized order to participants so that each participant evaluated each influencer once. Whether the disclosed or undisclosed influencer was presented to a participant was selected at random. As a result, we have a mixed study design of within-subjects and between-subjects comparisons. After having evaluated all 8 influencers, participants were thanked an received £3 to compensate for their time effort.

Measures. For each influencer, participants had to first state whether they are familiar with the presented influencer. After that came questions for measuring the level of mind attribution (taken from (Kozak et al., 2006)), followed by scales of perceived trust (adapted from D. Y. Kim and Kim (2021)), perceived uncanniness (adapted from Tinwell and Sloan (2014)), social presence (adapted from Gefen and Straub (1997)). The items for perceived trust and social presence are the same as employed on Study 1. In addition to our outcome variables, we added perceived influencer authenticity (Moulard et al., 2015), and perceived social and physical attractiveness of the influencer as control variables (H. Kim & Park, 2023). While both attractiveness as well as physical and social attractiveness are key influencing factors in social media perception in general (Batista & Chimenti, 2021; H. Kim & Park, 2023), perceived attractiveness might affect the emotional attachment with the influencer (H. Kim & Park, 2023). However, effects of authenticity and attractiveness might disappear when the influencer disclose whether it is virtual or human (Mirowska & Arsenyan, 2023). All questions were rated on a slider from 1 = totally disagree to 5 = totally agree. Finally, we asked participants to rate in how far they believed the influencer to be a human being as manipulation check (”The shown influencer is a human being.”, rated on 1 = Definitely not human to 5 = Definitely human).

7.3 Data Analysis

Analogous to Study 2, we conducted 2 analyses: one for testing the differences between human and virtual influencers for only the undisclosed posts, and one for the effect of disclosure for only the virtual influencers. For both analyses, we first ran mixed-effects models in which the conditions (i) human - virtual influencer; ii) disclosed - undisclosed virtual influencer) were set as fixed effects and each included construct was set a dependent variable. As random effects, we included individual differences between participants, as well as differences due to an interaction between influencer ethnicity and participant ethnicity. At this point it needs to be stated, that the interaction between influencer ethnicity and participant ethnicity did not reach significance for any of the included constructs.

In a second step, we wished to validate whether differences in mind attribution that were suggested by mPFC activation in Study 2 are predictors on uncanniness, social presence, and perceived trust. Therefore, we ran serial mediation analyses using Model 6 in the PROCESS macro for R by Hayes (2017) using 5,000 bootstrap samples and a confidence interval of 95 percent. Before running mediation analysis, we tested the construct reliability and validity. All constructs have sufficient reliability as signified by Cronbach's α ($.802 < \alpha < .955$) (Nunnally, 1978). Regarding convergent validity, the average variance extracted (AVE) for each construct exceeded the threshold of .50 for all constructs (Mind Attribution: $AVE = .679$, Uncanniness: $AVE > .506$, Social Presence: $AVE = .672$, Perceived Trust: $AVE = .601$). Further, the square root of the AVE for each construct exceeded inter-construct correlations ($Min_{SQRT(AVE)} = .711$; $Max_r = .678$), thereby providing discriminant validity between constructs (Fornell & Larcker, 1981).

7.4 Results

The manipulation check was significant for the perceived human-likeness of human versus virtual influencers ($F(1, 634.76) = 222.26, p < .001$). Further, we tested whether there are differences between the ethnic groups of included influencers. Results show that no significant differences were caused by influencer ethnicity ($F(3, 2.86) = 2.12, p = .283$), nor by random effects due to the ethnic group of the participants ($SD = .2194, p > .05$). Thereby, manipulation can be considered successful.

7.4.1 Differences in Virtual vs. Human Influencer Perception

We first tested whether there are significant differences between human and virtual influencers in the cases where no additional caption was provided that identified the influencer as human or as virtual. Overall our results show significant differences between human and virtual influencers (summarized in Figure 7). Mixed Effects models for each construct that considered participants and ethnicity as random effects showed a significant higher mind attribution for human influencers ($\beta = -0.806, 95\% CI[-0.9236, -0.6891], p < .001$), as well as higher social presence ($\beta = -0.587, 95\% CI[-0.6914, -0.4819], p < .001$), and higher perceived trust ($\beta = -0.217, 95\% CI[-0.3016, -0.1320], p < .001$). As hypothesized, perceived uncanniness of virtual influencers was higher than for human influencers ($\beta = 0.734, 95\% CI[0.6069, 0.8604], p < .001$). These effects remain significant when controlling for attractiveness and authenticity of the influencers.

Serial mediation analyses results show a significant predictive character of mind attribution on uncanniness, social presence, and perceived trust for the differences between

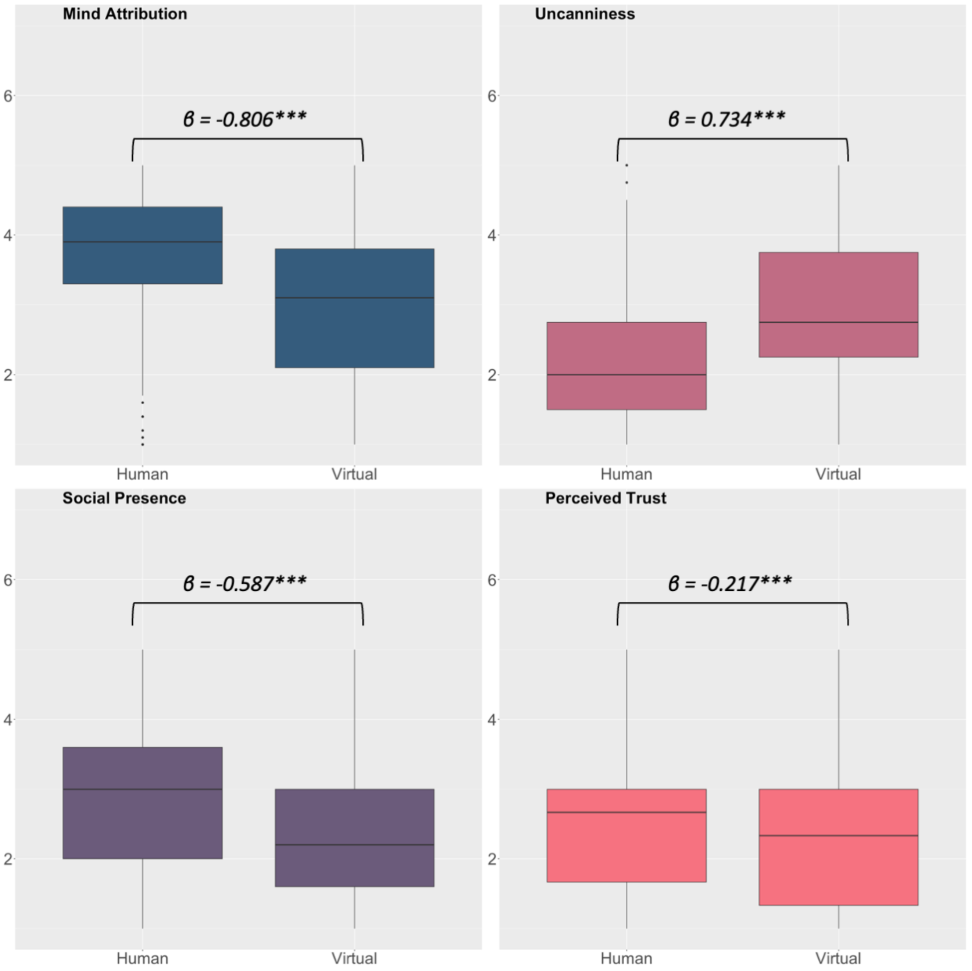


Figure 7: Results of Virtual vs. Human Influencers*Note.* $*p < .05$, $**p < .01$, $***p < .001$

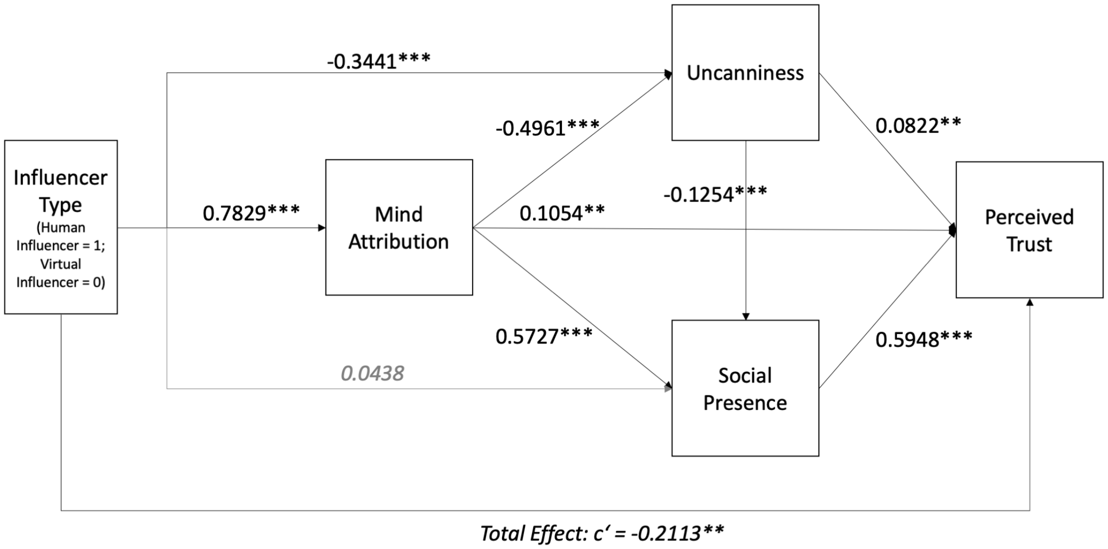


Figure 8: Serial Mediation Model Virtual vs. Human Influencers
Note. estimates are shown, $*p < .05$, $**p < .01$, $***p < .001$

human and virtual influencers (model depicted in Figure 8). It becomes evident that there is a significant positive mediation through mind attribution on perceived trust: influencer type → mind attribution → perceived trust ($\beta = 0.0525, 95\% CI[0.0322, 0.1386]$). Further, perceived trust is negatively mediated by uncanniness: influencer type → uncanniness → perceived trust ($\beta = -0.0283, 95\% CI[-0.0506, -0.0096]$). No mediating effects could be identified for social presence: influencer type → social presence → did not reach significance ($\beta = 0.026, 95\% CI[-0.0481, 0.102]$; because 0 is included in the confidence interval).

Further we identified significant negative serial mediation for influencer type → mind attribution → uncanniness → perceived trust ($\beta = -0.0319, 95\% CI[-0.0548, -0.0111]$). The mediation through influencer type → mind attribution → social presence → perceived trust was positive ($\beta = 0.2666, 95\% CI[0.2036, 0.3323]$). Both the serial mediation through influencer type → uncanniness → social presence → perceived trust ($\beta =$

$0.0257, 95\% CI[0.0097, 0.0462]$) and through influencer type → mind attribution → uncanniness → social presence → perceived trust ($\beta = 0.029, 95\% CI[0.0124, 0.0483]$) are positive. Given that the direct effect of influencer type on perceived trust reaches significance as well ($\beta = -0.1584, 95\% CI[-0.0548, -0.0111], p = .002$), we can speak at least of partial mediation that diminishes the negative effect of influencer type on perceived trust. In conclusion, the impact of influencer type (human vs. virtual influencer) on perceived trust is partially mediated by the levels of mind attribution and their impact on perceived uncanniness and social presence.

7.4.2 Perception and Processing of Disclosure in Influencers

Secondly, we tested whether there are significant differences between disclosed and undisclosed virtual influencers. In contrast to study 2, results show significant differences between disclosed and undisclosed virtual influencers (summarized in Figure 9). Mixed Effects models for each construct that considered participants and ethnicity as random effects showed a significant higher mind attribution for undisclosed compared to disclosed virtual influencers ($\beta = 0.479, 95\% CI[0.357, 0.6014], p < .001$), as well as higher social presence ($\beta = 0.2596, 95\% CI[0.157, 0.3626], p < .001$), and higher perceived trust ($\beta = 0.118, 95\% CI[0.0329, 0.203], p = .007$). Perceived uncanniness of disclosed virtual influencers was higher than for undisclosed virtual influencers ($\beta = -0.201, 95\% CI[-0.3288, -0.0726], p = .002$).

When including perceived attractiveness and authenticity of the virtual influencers as control variables, only mind attribution remains significantly higher for undisclosed virtual influencers. Serial mediation analyses results show a significant predictive character of mind attribution and social presence on perceived trust for the differences

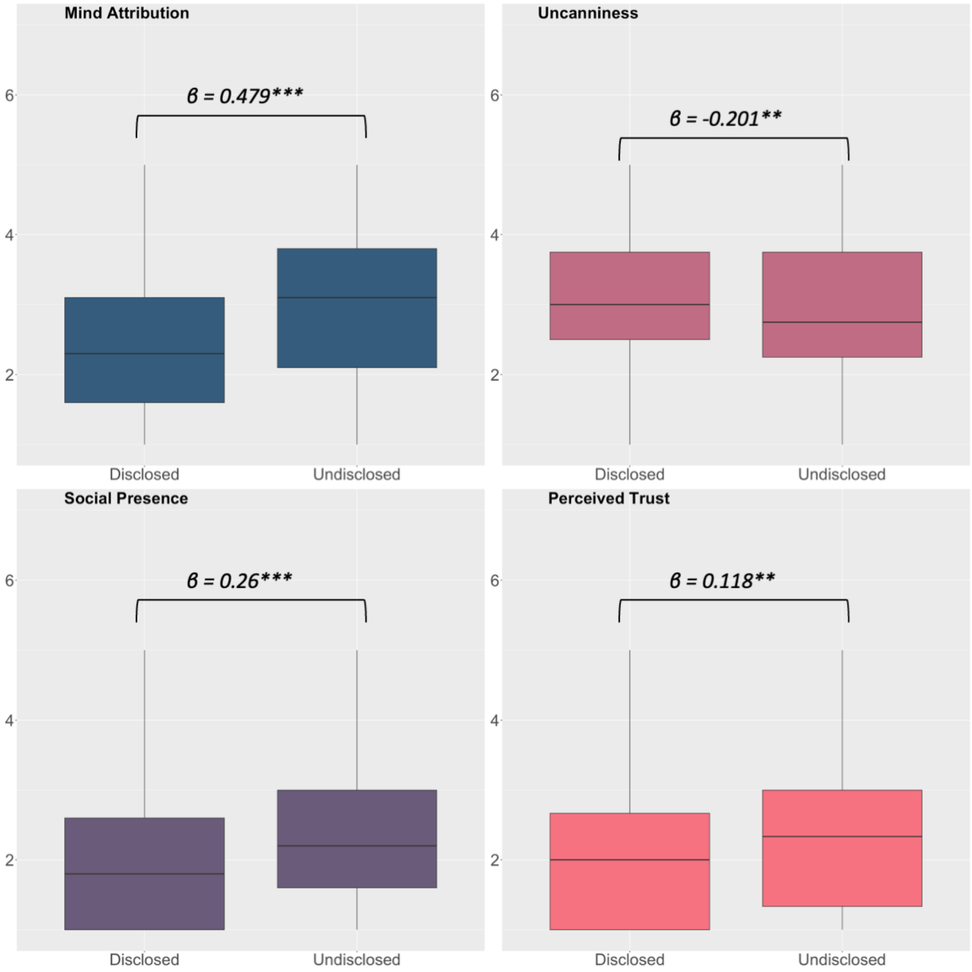


Figure 9: Results of Disclosed vs. Undisclosed Virtual Influencers
Note. $*p < .05$, $**p < .01$, $***p < .001$

between disclosed and undisclosed virtual influencers (model depicted in Figure 8). Given that the direct effect of disclosure on perceived trust did not reach significance ($\beta = 0.0548, 95\% CI[-0.0436, 0.1532], p = .275$), we find full mediation through the following paths.

It becomes evident that there is a significant negative mediation through mind attribution on perceived trust: influencer type → mind attribution → perceived trust ($\beta = -0.0607, 95\% CI[-0.1043, -0.0231]$). Further, perceived trust is serially mediated through mind attribution and social presence: influencer type → uncanniness → perceived trust ($\beta = -0.1694, 95\% CI[-0.2242, -0.1169]$), as well as through all included constructs: influencer type → mind attribution → uncanniness → social presence → perceived trust ($\beta = -0.0159, 95\% CI[-0.0269, -0.0073]$). All other mediations did not reach significance. It can be concluded that disclosure significantly impacts the level of mind attribution and perceived social presence as predictors on perceived trust.

8 Discussion

Although trust is often seen as a key indicator for successful social media content (Hess et al., 2009), our findings suggest that virtual influencers seem to be rated lower in trust compared to similar human influencers. This could be explained by a lower rating of perceived social presence and a higher perception of uncanniness towards virtual influencers. This relationship between trust and social presence as well as uncanniness has also been shown in previous research (Nissen & Jahn, 2021). Accordingly, photo-realistic human-like virtual influencers apparently have not yet crossed the uncanny

H	**Hypothesis**	**Result Study 2**	**Result Study 3**
1a	Human influencers will be rated higher in perceived trust when compared to similar virtual influencers.	Supported	Supported
1b	Human influencers will be rated higher in social presence when compared to similar virtual influencers.	Supported	Supported
1c	Human influencers will be rated lower in uncanniness when compared to similar virtual influencers.	Supported	Supported
2a	The perception of virtual influencers will lead to less mind attribution compared to the perception of similar human influencers.	–	Supported
2b	The perception of virtual influencers will lead to higher right mPFC activation compared to the perception of similar human influencers.	Supported	–
3a	Undisclosed virtual influencers will lead to lower perceived trust ratings than disclosed virtual influencers.	Rejected	Rejected
3b	Undisclosed virtual influencers will lead to lower perceived social presence ratings than disclosed virtual influencers.	Rejected	Rejected
3c	Undisclosed virtual influencers will lead to lower perceived uncanniness ratings than disclosed virtual influencers.	Supported	Supported
4a	Undisclosed virtual influencers will lead to higher mind attribution compared to disclosed virtual influencers.	–	Supported
4b	Undisclosed virtual influencers will lead to higher left mPFC activation compared to disclosed virtual influencers.	Supported	–

Table 5: Hypotheses Results

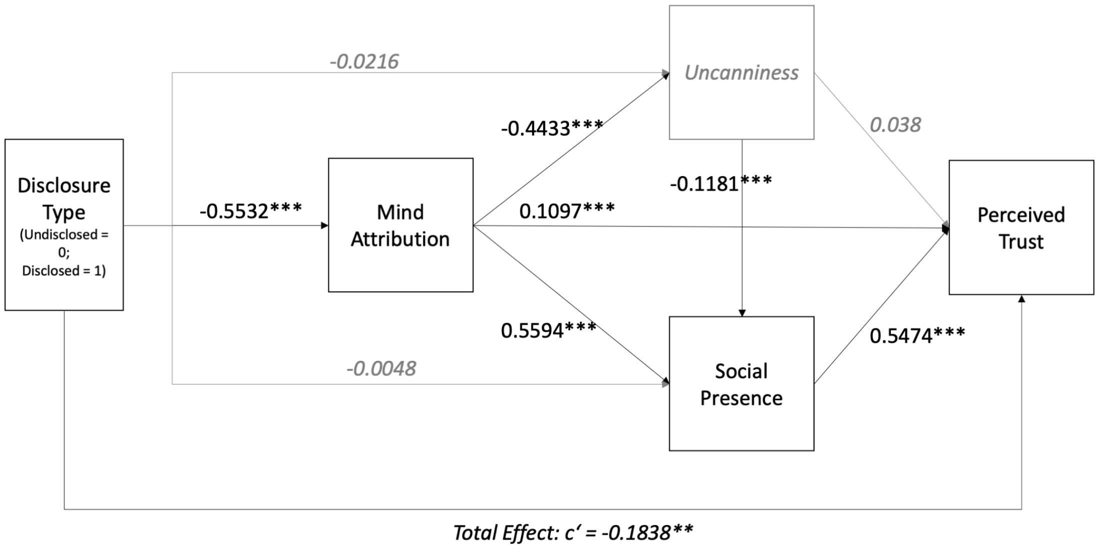


Figure 10: Serial Mediation Model Disclosed vs. Undisclosed Virtual Influencers
Note. *$p < .05$, **$p < .01$, ***$p < .001$

valley, as Seymour et al. (2021) assumed for virtual humans. Moreover, based on the significant mPFC activation identified in study 2, and in accordance with the findings of study 3, we are able to draw conclusions on the role of mind attribution toward virtual compared to human influencers. These are elaborated in the following.

8.1 Mind Attribution Explains Differences in Virtual Compared to Human Influencer Perception

Our findings suggest that virtual influencers are evaluated as less trustworthy, lesser socially present, and more uncanny than similar human influencers (H1a-c). Further considering our neural results, this perception of lower trust in virtual influencers seems to not always be an explicit cognitive-reflective process, but may be independent from whether people realize that an influencer is virtual or human (as shown in study 1).

This is in line with basic research on anthropomorphism suggesting that attributing human-like characteristics to non-human-like agents is a process which is independent from conscious opinions (Bhatti & Robert, 2023; Epley et al., 2007; Waytz et al., 2010). One explanation for this lower perceived trust might be a high level of uncertainty about to which degree a mind can be attributed. As became evident in our study 1 and other previous studies (e.g., Batista & Chimenti, 2021), participants were often unsure whether the presented influencer was human or virtual. Further supported by neuroscientific findings, this uncertainty of mind attribution seems to be reflected in elevated right mPFC activation for virtual influencers (H2b) which may signal negative valence, and lower ratings of mind attribution compared to human influencers (H2a).

Taken both the self-reported and neural results of study 2 and study 3 together, they may be an indicator for the interplay of self-rated trust, social presence, and uncanniness on the one hand, and mind attribution on the other. That is, the evaluation of virtual influencers can be separated into implicit social-perceptual automated processes (i.e., neural activity), and explicit slower cognitive-reflexive processes (i.e., self-reported data) (Keysers & Gazzola, 2007). Our results suggest that the perception of highly human-like virtual influencers results in an increased right mPFC activity due to them providing ambiguous information about their actual identity (Cornelius et al., 2023; Etkin et al., 2011). In line with this, from uncertainty about the degree to which a mind can be attributed to a virtual influencer, negative valence might arise as a result of an uncertainty or betrayal avoiding behavior, signified by increased right mPFC activity (Burgos-Robles et al., 2017; Hirshfield et al., 2019; Xue et al., 2009). In the slower cognitive-reflexive systems, the outcomes of these processes become evident in the results of all three included studies in the form of significantly lower perceived trust, social presence, and higher uncanniness.

Given that we identified such significant differences in the perception between human and virtual influencers in both the neural mechanisms and self-reported results, the question arises as to why accounts such as Lil' Miquela have such a large number of followers and attract a lot of attention (e.g., a high engagement rate). Related works that aim to explain why this happens have proposed that virtual influencers are also able to generate attractiveness and authenticity perceptions among social media users (Choudhry et al., 2022; H. Kim & Park, 2023; Lim & Lee, 2023). However, when controlling for attractiveness and authenticity perceptions, the constructs of mind attribution, uncanniness, social presence, and trust remain significant (see study 3). We also examined whether the ethnicity of the influencer in comparison to the ethnicity of the participants influences our results. However, our findings suggest that neither trust, nor perceived social presence, perceived uncanniness, or mind attribution are influenced by ethnicity. This was already indicated by Lim and Lee (2023) and reflects evidence that the role of trust, social presence, uncanniness, and mind attribution are valid regardless of the influencer's attractiveness, authenticity, and ethnicity.

Additionally, through our mediation model (Figure 8) we see that the influencer type significantly impacts the perceived uncanniness and mind attribution, both of which further explain perceived trust. Social presence, on the other hand, may not be a direct evaluation of the influencer type, but seems to be an outcome of mind attribution and of perceived uncanniness. Thereby, mind attribution and social presence as well as uncanniness and social presence seem to serially mediate the perceived trust. Therefore, mind attribution and uncanniness may be antecedent evaluation processes to social presence and trust. Both of these may also be reflected in the identified right mPFC activation in study 2, as this brain area evaluates another's intentions, emotions, cognition on the one hand (i.e., attribution of mind), but also encodes their value for the self (Frith & Frith, 2010; Satpute et al., 2014; Weaverdyck et al., 2021). As the right mPFC is more associated with negative valence (Davidson, 1998; Sackeim et al., 1978; Wager et al., 2003), this might a predecessor of the uncanniness ratings of the virtual influencers. This reasoning is supported by related works that have linked mPFC activation to uncanniness triggered by certain levels of human-likeness of artificial agents (Rosenthal-Von der Pütten et al., 2019; Wang & Quadflieg, 2014).

8.2 Disclosure Impacts Virtual Influencer Perceptions Through Mind Attribution

Previous research suggested disclosure as important influencing factor on the perception of social media influencers and their contents (Djurica & Mendling, 2020; Lim & Lee, 2023). Therefore, we further examined shifts of virtual influencer perception when they disclose themselves as non-human. While we found that self-disclosure did not result in an increase in perceived trust nor in social presence for virtual influencers in study 2, it did significantly impact perceived trust and social presence in study 3. The impact of disclosure on perceived uncanniness and mind attribution (or mPFC activation in study 2) seems consistent. However, when we controlled for the influencer's attractiveness and authenticity in study 3, disclosure only significantly impacted mind attribution while all other effects diminished.

This is further supported by the mediation model that shows how the impact of disclosure on trust is explained primarily by mind attribution, and in a second step by social presence through mind attribution (Figure 10). The changes in mind attribution as response to disclosure also significantly impacted perceived uncanniness; though this effect does not explain changes in perceived trust. Support for this effect is given by the results of study 2, where disclosed virtual influencers were rated significantly more uncanny than when they were undisclosed. In addition to this, undisclosed virtual influencers also lead to activation of the left mPFC.

As we have prior discussed the right mPFC as predecessing mechanism to mind attribution and uncanniness evaluations, a similar role may be ascribed to the left mPFC. On the neural level, undisclosed virtual influencers resulted in a higher left mPFC activation (H4b) than disclosed virtual influencers. While mPFC activation is in general related to mind attribution and evaluation processes, the left mPFC activation is usually a sign for more positive valence (Burgos-Robles et al., 2017; Hirshfield et al., 2019). Drawing back to the two-systems account of the ToM (Keysers & Gazzola, 2007; Meinhardt-Injac et al., 2018), the left mPFC seems to encode the differences in undisclosed versus disclosed virtual influencers in the implicit system (as supported in H4b). This activation assumes that a higher degree of mind attribution takes place when it is uncertain if the influencer is human or not, and that this is associated with more positive valence and, consequently, lower uncanniness. Both the level to which a mind is attributed and the perceived uncanniness become evident in self-report ratings which may point to evaluations of the reflective system. Upon disclosure, and therefore upon giving certainty that the influencer is non-human, the degree of mind attributed is decreased, and uncanniness perceptions increase. The increase in uncanniness may potentially be an outcome of the decision conflict between the mind that was first attributed, and the correction of this assumption after disclosure. Related works support this effect, arguing with an 'uncanny valley of mind attribution' that leads to rejection of too human-like identity of non-human agents (Stein & Ohler, 2017). In addition, we were able to deepen the findings of Lim and Lee (2023) by not only measuring perceived humanness toward virtual influencers, but also identifying mind attribution as an antecedent key process, and contextualizing it within the context of other relevant constructs.

Our theorizing is further in line with works in social media marketing context where the disclosure of an advertisement leads to a more negative impression of the influencer (Boerman et al., 2017; Karagür et al., 2022). Alibakhshi and Srivastava (2022) also suggested a mixed impact of self-disclosure of social media profile owners. However, this is contrary to the suggestions of Djurica and Mendling (2020) who proposed a positive correlation of trust in influencers and disclosure of advertisements in social media. Against our results, it can be said that disclosing a virtual influencer may reduce the level of mind attribution, which does not have positive consequences in the reflective attribution of trust, social presence, and uncanniness. More precisely, mind attribution is key to further evaluations of the influencer, which may include perceived attractiveness and authenticity as well, as the relevance of these constructs have been shown in several previous studies (Choudhry et al., 2022; H. Kim & Park, 2023; Lim & Lee, 2023). As a consequence, when disclosing virtual influencers in their posts, it needs to be done in a way that fosters mind attribution of the influencer, while transparently informing about its artificiality. This way, virtual influencers may be seen as ethically acceptable while still sparking curiosity and engagement in social media users.

8.3 Contribution to IS Research

We contribute to IS research by providing knowledge on human-like virtual influencer perception as an emerging technological phenomenon in research on human-computer interaction. By showing that mind attribution is a fundamental, antecedent process in the context of computer-generated actors in social media, we provide an explanation for perceptions of trust, uncanniness, and social presence. These perceptions are important factors for successful digital communication and interaction. As technological advances in many fields increase the complexity in the perception of human-like systems, mind attribution as an antecedent factor can serve to better understand the perception of other similar technologies such as conversational agents. In addition, our findings on mind attribution can contribute to the understanding and classification of the influence of other factors.

We found indication that trust is explained by levels of mind attribution, which is reduced for virtual influencers. Especially so when they disclose themselves as non-human agents. We conclude that trust seems to be an outcome of mind attribution rather than a key driver of success for virtual influencers as they were perceived lower in trust compared to human influencers. This theorizing also applies to disclosed versus undisclosed virtual influencers. Thereby, we provide a better understanding of trust in human-like technology and also contribute to IS research by examining the role of disclosure in social media. We revealed how self-disclosure does not directly affect trust and social presence, but reduced perceived uncanniness and mind attribution towards virtual influencers. For human-like artificial agents, it seems that low trust cannot be increased easily by transparency, since humans are deterred from trusting by uncertainty in mind attribution.

Moreover, our findings contribute to IS research by considering mind attribution from an IS perspective. By transferring mind attribution as advocated by the ToM to the context of trust in technology, we not only suggest how trust is not limited to an explicit perception, but also how it is related to a more implicit neural activation in the mPFC. Therefore, this might be an indicator for valence of mind attribution processes towards virtual influencers that result in reflective evaluations of uncanniness, social presence, and trust. These results are not only valuable for IS research on electronic media and cognitive research, but also benefit research on broader human-computer interaction and design research by suggesting a basis for research on similar human-like technologies.

With survey data and neuroimaging data, we combined self-reporting as a more common IS research method with fNIRS data as a direct measurement approach from neuroIS (Riedl et al., 2014b; Riedl & Léger, 2016). With this we provide an approach that allows IS scholars to examine implicit and explicit processes in a decision conflict when perceiving computer-generated content. Using both methods can provide knowledge on how slower cognitive-reflective processes relate to faster social-perceptual and more automated processes. In this way, we provide valuable knowledge that would not have been possible with traditional survey instruments or other self-report measurement methods alone.

8.4 Contribution to Practice

Our findings also contribute to practice in several ways. First, owners and creators of virtual influencers can learn that self-disclosure of the influencer does not necessarily impact the perception of the influencer. From a marketing point of view it therefore seems to be not necessary to label a virtual influencer as such.

Having identified mind attribution as the key driver to influencer perceptions, creators need to carefully craft the disclosing texts of virtual influencers in a way that supports mind attribution while still increasing transparency. That is, not disclosing virtual influencers at all would raise ethical issues with regard to deception of social media users. In the future, it needs to be decided as to whether there should be external regulations for the labeling of virtual influencers. Governments and policy makers should therefore discuss which computer-generated content should be labeled and what exactly the labeling should include in order to protect social media users and to identify the actors responsible for virtual influencers.

Second, organizations and their social media marketers can learn from our findings that people generally trust virtual influencers less than human influencers. This could also have a negative impact on the purchase intention of promoted products. Companies should therefore think carefully about what they want to achieve by cooperating with a virtual influencer. Prominent cases such as Lil' Miquela show that virtual influencers are still able to spark interest in social media users. Our findings suggest that designing for mind attribution may be key to successful marketing campaigns with virtual influencers.

Third, designers of virtual influencers can learn from our research. Although virtual influencers are sometimes not correctly recognized as such, effects that indicate a perceived uncanniness are still evident. Uncanniness is thereby not limited to the visual appearance of the virtual influencers but can also include autonomous and too randomized emotional parameters and real-time content. Equipping virtual influencers with more autonomy by using machine learning or natural language processing techniques might therefore cause even more uncanniness. Showing that a virtual influencer is rather scripted by a human team might reduce this perceived uncanniness of mind attribution. Therefore, designers should possibly continue the focus on trying to overcome the uncanny valley not only with photo-realism but also with other social cues to increase an implicit mind attribution by creating a unique identity and a suitable explanation about responsibility and content generation processes.

8.5 Limitations and Future Research

This study comes with some limitations that provide opportunities for future research. First, we presented the influencers in a short time period in study 2, which could have impacted the perceived trust and the impact of disclosure. In study 3, this shortcoming is partly overcome. However, we still considered only individual posts and not the storytelling that could be realized through a number of posts of an influencer. Future studies are therefore advised to conduct long-term projects in order to check whether trust in virtual influencers and mind attribution increases over time. Furthermore, it may be reasonable to target disclosure in more detail and address different cues that could be given through disclosure that may further enhance rather than decrease mind attribution of the influencer.

Second, fNIRS studies require a high complexity in the experimental setting resulting in a usually smaller sample size. Future studies therefore need to further validate our findings with more participants and different virtual influencers. Other direct measurement approaches such as EEG could also be used in future studies for further validation. Third, our study was limited to images of existing virtual and human influencers in Instagram to ensure comparability of the influencers' content. Even though we chose images that looked as similar as possible for the comparisons, slightly different poses and gestures may have influenced our results (Kitamura & Watanabe, 2023) or the style of the picture (Qiu et al., 2015). However, since we did not find any differences between the various influencer comparisons, especially in study 3, this influence is rather unlikely. Andrade et al. (2015) stated that images are an important source of data for IS research, future studies could consider the perception of text or video content and the impact of the social media platform that the content is posted on. As an example, social presence and mind attribution might be higher for video content due to more visible social cues. As virtual influencers are not limited to human-like influencers, it would also be promising to further compare trust in human-like virtual influencers with non-human like influencers. According to the two-systems account of the ToM (Meinhardt-Injac et al., 2018), it is likely that people attribute a higher level of mind to more human-like virtual influencers due to the automated and implicit evaluation of their faces and facial expressions (Willis & Todorov, 2006).

IS research, with its knowledge of ethical principles of socio-technical systems and characteristics of new technologies, provides an excellent foundation for further research into the necessary regulations for labeling computer-generated content from characters such as virtual influencers. Future research should take advantage of this interdisciplinary nature of IS to explore how such content can be labeled. This could lead not only to guidelines for policy makers, but also to advice for organizations on how to use virtual influencers in a way that is both consistent with ethical norms and useful.

9 Conclusion

Although people cannot always determine whether an influencer is human or not, our findings suggest that perceptions of virtual influencers are more negative than for human influencers. We identified that the level of mind attribution to such influencers is the key antecedent to evaluations of uncanniness, social presence, and trust. At the same time, mind attribution is influenced neither by ethnicity nor by perceived attractiveness or authenticity. Disclosing of virtual influencers as such decreases levels of mind attribution, resulting in higher uncanniness and lower perceived mind attribution. It becomes evident, that mind attributing processes of virtual influencers take place in the mPFC and seem to be encoded with their value (left hemisphere = positive, right hemisphere = negative) for the self. The results of this coding can be seen in self-report perceptions of mind attribution, as well as uncanniness, social presence, and trust. We therefore conclude that designing virtual influencers and their posts on social media in a way that they foster positive mind attribution is a key antecedent to their success.

References

Ahn, R. J., Cho, S. Y., & Tsai, W. S. (2022). Demystifying Computer-Generated Imagery (CGI) Influencers : The Effect of Perceived Anthropomorphism and Social Presence on Brand Outcomes. *Journal of Interactive Advertising*, *0*(0), 1–9.

Alibakhshi, R., & Srivastava, S. C. (2022). Post-story: Influence of introducing story feature on social media posts. *Journal of Management Information Systems*, *39*(2), 573–601.

Andrade, A. D., Urquhart, C., & Arthanari, T. S. (2015). Seeing for understanding: Unlocking the potential of visual research in information systems. *Journal of the Association for Information Systems*, *16*(8), 646–673.

Arsenyan, J., & Mirowska, A. (2021). Almost human? A comparative case study on the social media presence of virtual influencers. *International Journal of Human Computer Studies*, *155*(June), 102694.

Baier, A. (1986). Trust and antitrust. *ethics*, *96*(2), 231–260.

Barker, J. W., Aarabi, A., & Huppert, T. J. (2013). Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS. *Biomedical Optics Express*, *4*(8), 1366.

Batista, A., & Chimenti, P. (2021). " Humanized Robots ": A Proposition of Categories to Understand Virtual Influencers. *Australasian Journal of Information Systems*, *25*, 1–27.

Benbasat, I., & Wang, W. (2005). Trust In and Adoption of Online Recommendation Agents. *Journal of the Association for Information Systems*, *6*(3), 72–101.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Bente, G., Rüggenberg, S., Krämer, N. C., & Eschenburg, F. (2008). Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations. *Human Communication Research*, *34*(2), 287–318.

Bhatti, S. C., & Robert, L. P. (2023). What Does It Mean to Anthropomorphize Robots? Food For Thought for HRI Research. *ACM/IEEE International Conference on Human-Robot Interaction*, 422–425.

Boerman, S. C., Willemsen, L. M., & Van Der Aa, E. P. (2017). "This Post Is Sponsored": Effects of Sponsorship Disclosure on Persuasion Knowledge and Electronic Word of Mouth in the Context of Facebook. *Journal of Interactive Marketing*, *38*, 82–92.

Brigadoi, S., Ceccherini, L., Cutini, S., Scarpa, F., Scatturin, P., Selb, J., Gagnon, L., Boas, D. A., & Cooper, R. J. (2014). Motion artifacts in functional near-infrared spectroscopy: A comparison of motion correction techniques applied to real cognitive data. *NeuroImage*, *85*, 181–191.

Burgos-Robles, A., Kimchi, E. Y., Izadmehr, E. M., Porzenheim, M. J., Ramos-Guasp, W. A., Nieh, E. H., Felix-Ortiz, A. C., Namburi, P., Leppla, C. A., Presbrey, K. N., Anandalingam, K. K., Pagan-Rivera, P. A., Anahtar, M., Beyeler, A., & Tye, K. M. (2017). Amygdala inputs to prefrontal cortex guide behavior amid conflicting cues of reward and punishment. *Nature Neuroscience*, *20*(6), 824–835.

Choudhry, A., Han, J., Xu, X., & Huang, Y. (2022). "i Felt a Little Crazy Following a 'Doll'". *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP), 1–28.

Cinciute, S., Daktariunas, A., & Ruksenas, O. (2018). Hemodynamic effects of sex and handedness on the Wisconsin Card Sorting Test: the contradiction between neuroimaging and behavioural results. *PeerJ*, *6*, e5890.

Cornelius, S., & Leidner, D. (2021). Acceptance of anthropomorphic technology: A literature review. *2020-January*(1), 6422–6431.

Cornelius, S., Leidner, D. E., & Benbya, H. (2023). Credibility of Virtual Influencers: The Role of Design Stimuli, Knowledge Cues, and User Disposition. *Proceedings of the Annual Hawaii International Conference on System Sciences (HICSS)*, 3401–3410.

Dabiran, E., Wang, F., & Farivar, S. (2022). Virtual Influencer Marketing: Anthropomorphism and Its Effect. *Proceedings in: The European Conference on Information Systems*, Timisoara, Romania.

Davidson, R. J. (1992). Anterior Cerebral Asymmetry and the Nature of Emotion. *Brain and Cognition*, *151*, 125–151.

Davidson, R. J. (1998). Affective Style and Affective Disorders: Perspectives from Affective Neuroscience. *Cognition and Emotion*, *12*(3), 307–330.

Delpy, D. T., Cope, M., van der Zee, P., Arridge, S., Wray, S., & Wyatt, J. (1988). Estimation of optical pathlength through tissue from direct time of flight measurement. *Physics in Medicine and Biology*, *33*(12), 1433–1442.

DiYanni, C., Nini, D., Rheel, W., & Livelli, A. (2012). 'I Won't Trust You if I Think You're Trying to Deceive Me': Relations Between Selective Trust, Theory of Mind, and Imitation in Early Childhood. *Journal of Cognition and Development*, *13*(3), 354–371.

Djurica, D., & Mendling, J. (2020). Impact of Influencer Type and Advertisement Disclosure on Perceived Trust, Credibility and Purchase Intention. *Proceedings in: The American Conference on Information Systems*, *15*(Virtual Conference). https://aisel.aisnet.org/amcis2020/social_computing/social_computing/15

Dunn, J. (2000). Trust and Political Agency. In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations* (electronic, pp. 73–93). Department of Sociology, University of Oxford.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886.

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences*, *15*(2), 85–93.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, *132*, 138–161.

Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage*, *63*(2), 921–935
ICIS Paper.

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39.

Frith, U., & Frith, C. (2010). The social brain: Allowing humans to boldly go where no other species has been. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1537), 165–175.

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication*, *1*(1), 71–85.

Gefen, D., & Straub, D. W. (1997). Gender differences in the perception and use of e-mail: An extension to the technology acceptance model. *MIS Quarterly*, *21*(4), 389–400. Retrieved October 17, 2022, from http://www.jstor.org/stable/249720

Geller, T. (2008). Overcoming the Uncanny Valley. *IEEE Computer Graphics and Applications*, *28*(4), 11–17.

Goodwin, J. R., Gaudet, C. R., & Berger, A. J. (2014). Short-channel functional near-infrared spectroscopy regressions improve when source-detector separation is reduced. *Neurophotonics*, *1*(1), 015002.

Hanson, D., Olney, A., Pereira, I. A., & Zielke, M. (2005). Upending the uncanny valley. *AAAI Workshop - Technical Report*, *WS05-11*(July), 24–31.

Hayes, A. F. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis, Second Edition : A Regression-Based Approach*. Guilford Press.

Hess, T., Fuller, M., & Campbell, D. (2009). Designing interfaces with social presence: Using vividness and extraversion to create social recommendation agents. *Journal of the Association for Information Systems*, *10*(12), 889–919.

Hirshfield, L., Bobko, P., Barelka, A., Sommer, N., & Velipasalar, S. (2019). Toward Interfaces that Help Users Identify Misinformation Online: Using fNIRS to Measure Suspicion. *Augmented Human Research*, *4*(1), 1–13.

Hofeditz, L., Ehnis, C., Bunker, D., Brachten, F., & Stieglitz, S. (2019). Meaningful Use Of Social Bots? Possible Applications In Crisis Communication During Disasters. *European Conference on Information Systems*.

Hofeditz, L., Erle, L., & Timm, L. (2023). How Virtuous are Virtual Influencers ? – A Qualitative Analysis of Virtual Actors ’ Virtues on Instagram. *Hawaii International Conference on System Sciences*.

Holtgraves, T., & Han, T. L. (2007). A procedure for studying online conversational processing using a chat bot. *Behavior Research Methods*, *39*(1), 156–163.

Hoshi, Y., Kobayashi, N., & Tamura, M. (2001). Interpretation of near-infrared spectroscopy signals: a study with a newly developed perfused rat brain model. *Journal of Applied Physiology*, *90*(5), 1657–1662.

Huppert, T. (2016). Commentary on the statistical properties of noise and its implication on general linear models in functional near-infrared spectroscopy. *Neurophotonics*, *3*(1), 1–10.

Huppert, T., Hoge, R., Diamond, S., Franceschini, M., & Boas, D. (2006). A temporal comparison of BOLD, ASL, and NIRS hemodynamic responses to motor stimuli in adult humans. *NeuroImage*, *29*(2), 368–382.

Karagür, Z., Becker, J.-M., Klein, K., & Edeling, A. (2022). How, why, and when disclosure type matters for influencer marketing. *International Journal of Research in Marketing*, *39*(2), 313–335.

Kennedy, M. S., Ferrell, L. K., & LeClair, D. T. (2001). Consumers' trust of salesperson and manufacturer: An empirical study [International Marketing Strategy]. *Journal of Business Research*, *51*(1), 73–86.

Keysers, C., & Gazzola, V. (2007). Spatial cognition in apes and humans. *Trends in Cognitive Sciences*, *11*(5), 192–194.

Killgore, W. D. S., & Yurgelun-Todd, D. A. (2007). The right-hemisphere and valence hypotheses: could they both be right (and sometimes left)? *Social Cognitive and Affective Neuroscience*, *2*(3), 240–250.

Kim, D. Y., & Kim, H. Y. (2021). Trust me, trust me not: A nuanced view of influencer marketing on social media. *Journal of Business Research*, *134*, 223–232.

Kim, H., & Park, M. (2023). Virtual influencers' attractiveness effect on purchase intention: A moderated mediation model of the Product–Endorser fit with the brand. *Computers in Human Behavior*, *143*(107703).

Kitamura, M., & Watanabe, K. (2023). Managed postures modulate social impressions after limited and unlimited time exposure. *Current Psychology*, *42*(5), 3957–3967.

Kocsis, L., Herman, P., & Eke, A. (2006). The modified Beer–Lambert law revisited. *Physics in Medicine and Biology*, *51*(5), N91–N98.

Kozak, M. N., Marsh, A. A., & Wegner, D. M. (2006). What do i think you're doing? Action identification and mind attribution. *Journal of Personality and Social Psychology*, *90*(4), 543–555.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS ONE*, *3*(7).

Krampe, C., Gier, N., & Kenning, P. (2017). Beyond traditional neuroimaging: Can mobile fNIRS add to NeuroIS? In F. D. Davis, R. Riedl, J. vom Brocke, P.-M. Léger, A. Randolph, & T. Fischer (Eds.), *Lecture notes in information systems and organisation* (pp. 151–157). Springer Nature
ICIS.

Kretzer, M., & Maedche, A. (2018). Designing social nudges for enterprise recommendation agents: An investigation in the business intelligence systems context. *Journal of the Association for Information Systems*, *19*(12), 1145–1186.

Li, Y., Chen, R., Zhang, S., Turel, O., Bechara, A., Feng, T., Chen, H., & He, Q. (2019). Hemispheric mPFC asymmetry in decision making under ambiguity and risk: An fNIRS study. *Behavioural Brain Research*, *359*(June 2018), 657–663.

Lieberman, M. D. (2007). Social Cognitive Neuroscience: A Review of Core Processes. *Annual Review of Psychology*, *58*(1), 259–289
Diss.

Lim, R. E., & Lee, S. Y. (2023). "You are a virtual influencer!": Understanding the impact of origin disclosure and emotional narratives on parasocial relationships and virtual influencer credibility. *Computers in Human Behavior*, *148*(June), 107897.

Lu, B., Fan, W., & Zhou, M. (2016). Social presence, trust, and social commerce purchase intention: An empirical research. *Computers in Human Behavior*, *56*, 225–237.

Luck, S. J. (2014). *An Intorduction to the Event-Related Potential Technique* (2nd ed.). MIT Press.

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. T. (2009). Too real for comfort? uncanny responses to computer generated faces [Including the Special Issue: Enabling elderly users to create and share self authored multimedia content]. *Computers in Human Behavior*, *25*(3), 695–710.

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, *146*, 22–32.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, *2*(2).

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359.

Meinhardt-Injac, B., Daum, M. M., Meinhardt, G., & Persike, M. (2018). The two-systems account of theory of mind: Testing the links to social-perceptual and cognitive abilities. *Frontiers in Human Neuroscience*, *12*(January), 1–12.

Mirowska, A., & Arsenyan, J. (2023). Sweet escape: The role of empathy in social media engagement with human versus virtual influencers. *International Journal of Human Computer Studies*, *174*(February).

Miura, N., Sugiura, M., Takahashi, M., Miyamoto, A., & Kawashima, R. (2009). The effect of emotional valence and body structure on emotional empathy to humanoid robot: an fMRI study. *NeuroImage*, *47*(Supplement 1), S39–S41.

Molenaar, K. (2022). Discover The Top 12 Virtual Influencers for 2023 – Listed and Ranked! https://influencermarketinghub.com/virtual-influencers/#toc-3

Molenberghs, P., & Morrison, S. (2014). The role of the medial prefrontal cortex in social categorization. *Social Cognitive and Affective Neuroscience*, *9*(3), 292–296.

Mori, M. (1970). The uncanny valley. *Energy*, *7*(4), 33–35.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley. *IEEE Robotics and Automation Magazine*, *19*(2), 98–100.

Mou, W., Ruocco, M., Zanatto, D., & Cangelosi, A. (2020). When Would You Trust a Robot? A Study on Trust and Theory of Mind in Human-Robot Interactions. *29th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2020*, 956–962.

Moulard, J. G., Garrity, C. P., & Rice, D. H. (2015). What Makes a Human Brand Authentic? Identifying the Antecedents of Celebrity Authenticity. *Psychology & Marketing*, *32*(2), 173–186.

Mouritzen, S. L. T., Penttinen, V., & Pedersen, S. (2023). Virtual influencer marketing: the good, the bad and the unreal. *European Journal of Marketing*.

Moustakas, E., Lamba, N., Mahmoud, D., & Ranganathan, C. (2020). Blurring lines between fiction and reality: Perspectives of experts on marketing effectiveness of virtual influencers. *International Conference on Cyber Security and Protection of Digital Services, Cyber Security 2020*.

Mustafa, M., Guthe, S., Tauscher, J. P., Goesele, M., & Magnor, M. (2017). How human am I? EEG-based evaluation of animated virtual characters. *Conference on Human Factors in Computing Systems - Proceedings*, *2017-May*, 5098–5108.

Mustafa, M., & Magnor, M. (2016). EEG based analysis of the perception of computer-generated faces. *ACM International Conference Proceeding Series*.

Nass, C., Steuer, J., & Tauber, E. R. (1994). are Social Actors. *Human Factors*, 72–78.

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance.

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(8), 2–4.

Nissen, A. (2020). Psychological and Physiological Effects of Color Use on eCommerce Websites: a Neural Study Using fNIRS. *International Conference on Information Systems (ICIS)*.

Nissen, A., Conrad, C., & Newman, A. (2023). Are You Human? Investigating the Perceptions and Evaluations of Virtual Versus Human Instagram Influencers. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*.

Nissen, A., & Jahn, K. (2021). Between anthropomorphism, trust, and the uncanny valley: A dual-processing perspective on perceived trustworthiness and its mediating effects on use intentions of social robots. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 360–369.

Nissen, A., & Krampe, C. (2021). Why He Buys It and She Doesn't – Exploring Self-Reported and Neural Gender Differences in the Perception of eCommerce Websites. *Computers in Human Behavior*, *121*(April).

Noah, J. A., Ono, Y., Nomoto, Y., Shimada, S., Tachibana, A., Zhang, X., Bronner, S., & Hirsch, J. (2015). fMRI Validation of fNIRS Measurements During a Naturalistic Task. *Journal of Visualized Experiments*, (100), 5–9.

Nowak, K. L., & Fox, J. (2018). Avatars and computer-mediated communication: A review of the definitions, uses, and effects of digital representations. *Review of Communication Research*, *6*, 30–53.

Nunnally, J. (1978). *Psychometric Theory* (2nd ed.). McGraw-Hill.

Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., & Boyle, M. P. (2020). The Ineffectiveness of Fact-Checking Labels on News Memes and Articles. *Mass Communication and Society*, *23*(5), 682–704.

Ogonowski, A., Montandon, A., Botha, E., & Reyneke, M. (2014). Should new online stores invest in social presence elements? The effect of social presence on initial trust formation. *Journal of Retailing and Consumer Services*, *21*(4), 482–491.

Ozdemir, O., Kolfal, B., Messinger, P. R., & Rizvi, S. (2023). Human or virtual: How influencer type shapes brand attitudes. *Computers in Human Behavior*, *145*(February), 107771.

Park, G., Nan, D., Park, E., Kim, K. J., Han, J., & Del Pobil, A. P. (2021). Computers as Social Actors? Examining How Users Perceive and Interact with Virtual Influencers on Social Media. *Proceedings of the 2021 15th International Conference on Ubiquitous Information Management and Communication, IMCOM 2021*, 10–15.

Pavlou, P. A., Liang, H., & Xue, Y. (2007). Understanding and mitigating uncertainty in online exchange relationships: A principal-agent perspective. *MIS Quarterly*, *31*(1), 105–136. Retrieved October 10, 2022, from http://www.jstor.org/stable/25148783

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2020). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1464*(1), 5–29.

Qiu, L., Lu, J., Yang, S., Qu, W., & Zhu, T. (2015). What does your selfie say about you? *Computers in Human Behavior*, *52*, 443–449.

Riedl, R., Davis, F. D., & Hevner, A. R. (2014b). Towards a neuroIS research methodology: Intensifying the discussion on Methods, Tools, And Measurement. *Journal of the Association for Information Systems*, *15*(10), 1–35.

Riedl, R., & Léger, P.-M. (2016). *Tools in neurois research: An overview*. Springer Berlin, Heidelberg.

Riedl, R., Mohr, P., Kenning, P., Davis, F., & Heekeren, H. (2014a). Trusting humans and avatars: A brain imaging study based on evolution theory. *Journal of Management Information Systems*, *30*(4), 83–114.

Robinson, B. (2020). Towards an ontology and ethics of virtual influencers. *Australasian Journal of Information Systems*, *24*, 1–8.

Rosenthal-Von der Pütten, A. M., Krämer, N. C., Maderwald, S., Brand, M., & Grabenhorst, F. (2019). Neural mechanisms for accepting and rejecting artificial social partners in the uncanny valley. *Journal of Neuroscience*, *39*(33), 6555–6570.

Ross, B., Heisel, J., Jung, A. K., & Stieglitz, S. (2018). Fake news on social media: The (in)effectiveness of warning messages. *International Conference on Information Systems 2018, ICIS 2018*, 1–17.

Rousseau, D. M., Sitkin, S., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross discipline view of trust. *Academy of Management Review*, *23*(3), 393–404.

Ruocco, M., Mou, W., Cangelosi, A., Jay, C., & Zanatto, D. (2021). Theory of mind improves human's trust in an iterative human-robot game. *HAI 2021 - Proceedings of the 9th International User Modeling, Adaptation and Personalization Human-Agent Interaction*, 227–234.

Saager, R. B., & Berger, A. J. (2005). Direct characterization and removal of interfering absorption trends in two-layer turbid media. *Journal of the Optical Society of America A*, *22*(9), 1874.

Sackeim, H. A., Gur, R. C., & Saucy, M. C. (1978). Emotions are expressed more intensely on the left side of the face. *Science*, *202*(4366), 434–436.

Salmaso, D., & Longoni, A. M. (1985). Problems in the Assessment of Hand Preference. *Cortex*, *21*(4), 533–549.

Santosa, H., Zhai, X., Fishburn, F., & Huppert, T. (2018). The NIRS Brain AnalyzIR toolbox. *Algorithms*, *11*(5).

Satpute, A. B., Badre, D., & Ochsner, K. N. (2014). Distinct regions of prefrontal cortex are associated with the controlled retrieval and selection of social information. *Cerebral Cortex*, *24*(5), 1269–1277.

Saxe, R., & Baron-Cohen, S. (2006). The neuroscience of theory of mind. *Social neuroscience*, *1*(3-4), 1–9.

Schmitz, J., Lor, S., Klose, R., Güntürkün, O., & Ocklenburg, S. (2017). The functional genetics of handedness and language lateralization: Insights from gene ontology, pathway and disease association analyses. *Frontiers in Psychology*, *8*(JUL), 1–12.

Scholkmann, F., Kleiser, S., Metz, A. J., Zimmermann, R., Mata Pavia, J., Wolf, U., & Wolf, M. (2014). A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *NeuroImage*, *85*, 6–27.

Seymour, M., Yuan, L., Dennis, A. R., & Riemer, K. (2021). Have we crossed the uncanny valley? Understanding affinity, trustworthiness, and preference for realistic digital humans in immersive environments. *Journal of the Association for Information Systems*, *22*(3), 591–617.

Shareef, M. A., Kapoor, K. K., Mukerji, B., Dwivedi, R., & Dwivedi, Y. K. (2020). Group behavior in social media: Antecedents of initial trust formation. *Computers in Human Behavior*, *105*, 106225.

Shevlin, M., Walker, S., Davies, M. N. O., Banyard, P., & Lewis, C. A. (2003). Can you judge a book by its cover? Evidence of self–stranger agreement on personality at zero acquaintance. *Personality and Individual Differences*, *35*(6), 1373–1383.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. John Wiley Sons Ltd.

Simpson, T. W. (2012). What is trust? *Pacific Philosophical Quarterly*, *93*(4), 550–569.

Skjuve, M., Haugstveit, I. M., Følstad, A., & Brandtzaeg, P. B. (2019). Help! Is my chatbot falling into the uncanny valley? An empirical study of user experience in human-chatbot interaction. *Human Technology*, *15*(1), 30–54.

Song, C., & Lee, J. (2016). Citizens' use of social media in government, perceived transparency, and trust in government. *Public Performance & Management Review*, *39*(2), 430–453.

Srivastava, S. C., & Chandra, S. (2018). Social presence in virtual world collaboration: An uncertainty reduction perspective using a mixed methods approach1. *MIS Quarterly: Management Information Systems*, *42*(3), 779–803.

Stein, J.-P., & Ohler, P. (2017). Venturing into the uncanny valley of mind—the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition*, *160*, 43–50.

Strangman, G., Culver, J. P., Thompson, J. H., & Boas, D. A. (2002). A Quantitative Comparison of Simultaneous BOLD fMRI and NIRS Recordings during Functional Brain Activation. *NeuroImage*, *17*(2), 719–731.

Times. (2018). The 25 Most Influential People on the Internet. https://time.com/5324130/most-influential-internet/

Tinwell, A., & Sloan, R. J. (2014). Children's perception of uncanny human-like virtual characters. *Computers in Human Behavior*, *36*(July 2014), 286–296.

Toronov, V., Walker, S., Gupta, R., Choi, J. H., Gratton, E., Hueber, D., & Webb, A. (2003). The roles of changes in deoxyhemoglobin concentration and regional cerebral blood volume in the fMRI BOLD signal. *NeuroImage*, *19*(4), 1521–1531.

Venkatesh, V., Thong, J. Y., Chan, F. K., & Hu, P. J. (2016). Managing citizens' uncertainty in e-government services: The mediating and moderating roles of transparency and trust. *Information Systems Research*, *27*(1), 87–111.

Wager, T. D., Phan, K. L., Liberzon, I., & Taylor, S. F. (2003). Valence, gender, and lateralization of functional brain anatomy in emotion: A meta-analysis of findings from neuroimaging. *NeuroImage*, *19*(3), 513–531.

Wang, Y., & Quadflieg, S. (2014). In our own image? Emotional and neural processing differences when observing human-human vs human-robot interactions. *Social Cognitive and Affective Neuroscience*, *10*(11), 1515–1524.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117.

Weaverdyck, M. E., Thornton, M. A., & Tamir, D. I. (2021). The representational structure of mental states generalizes across target people and stimulus modalities. *NeuroImage*, *238*(December 2020), 118258.

Wijeakumar, S., Huppert, T. J., Magnotta, V. A., Buss, A. T., & Spencer, J. P. (2017). Validating an image-based fNIRS approach with fMRI and a working memory task. *NeuroImage*, *147*(December 2015), 204–218.

Willis, J., & Todorov, A. (2006). First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face.

Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, *5*(3), 277–283.

Xue, G., Lu, Z., Levin, I. P., Weller, J. A., Li, X., & Bechara, A. (2009). Functional dissociations of risk and reward processing in the medial prefrontal cortex. *Cerebral Cortex*, *19*(5), 1019–1027.

Yücel, M. A., Selb, J., Aasted, C. M., Lin, P.-Y., Borsook, D., Becerra, L., & Boas, D. A. (2016). Mayer waves reduce the accuracy of estimated hemodynamic response functions in functional near-infrared spectroscopy. *Biomedical Optics Express*, *7*(8), 3078.

Zhou, L., & Xue, F. (2021). Show products or show people: an eye-tracking study of visual branding strategy on Instagram. *Journal of Research in Interactive Marketing*, *15*(4), 729–749.

Paper 7: Do You Trust an AI-Journalist? A Credibility Analysis of News Content With AI-Authorship

Type (Ranking, Impact Factor)	Conference article (B, N/A)
Status	Published
Rights and permissions	Open Access
Authors	Hofeditz, L., Mirbabaie, Mi., Stieglitz, S., & Holstein, J.
Year	2021
Outlet	European Conference on Information Systems
Permalink / DOI	https://aisel.aisnet.org/ecis2021_rp/50
Full citation	Hofeditz, L., Mirbabaie, Mi., Stieglitz, S., & Holstein, J. (2021). Do You Trust An AI-Journalist? A Credibility Analysis Of News Content With AI-Authorship. *European Conference on Information Systems.* https://aisel.aisnet.org/ecis2021_rp/50.

Do you Trust an AI-Journalist? A Credibility Analysis of News Content with AI-Authorship

Conference Paper · June 2021

4 authors, including:

Lennart Hofeditz
Universität Potsdam
24 PUBLICATIONS **178** CITATIONS
SEE PROFILE

Milad Mirbabaie
Universität Paderborn
132 PUBLICATIONS **1,838** CITATIONS
SEE PROFILE

Stefan Stieglitz
Universität Potsdam
320 PUBLICATIONS **7,494** CITATIONS
SEE PROFILE

6-14-2021

DO YOU TRUST AN AI-JOURNALIST? A CREDIBILITY ANALYSIS OF NEWS CONTENT WITH AI-AUTHORSHIP

Lennart Hofeditz
University of Duisburg-Essen, lennart.hofeditz@uni-due.de

Milad Mirbabaie
Paderborn University, milad.mirbabaie@uni-paderborn.de

Jasmin Holstein
University of Duisburg-Essen, jasmin.holstein@stud.uni-due.de

Stefan Stieglitz
Universität Duisburg-Essen, stefan.stieglitz@uni-due.de

DO YOU TRUST AN AI-JOURNALIST? A CREDIBILITY ANALYSIS OF NEWS CONTENT WITH AI-AUTHORSHIP

Research Paper

Lennart Hofeditz, University of Duisburg-Essen, Duisburg, Germany, lennart.hofeditz@uni-due.de

Milad Mirbabaie, Paderborn University, Paderborn, Germany, milad.mirbabaie@uni-paderborn.de

Stefan Stieglitz, University of Duisburg-Essen, , Duisburg, Germany, stefan.stieglitz@uni-due.de

Jasmin Holstein, University of Duisburg-Essen, Duisburg, Germany, holstein.jasmin@stud.uni-due.de

Abstract

Due to the increasing amount of data, media companies increasingly use algorithms and artificial intelligence (AI) for both researching topics and autonomously creating journalistic articles and for publishing content on social media. However, previous studies found an increase of uncertainty of social media as a source of information, due to fake news and nontransparency of the source. Research on the credibility of using AI in journalism and trust in AI-generated articles is still in early stages. Therefore, we conducted an online survey (n = 122) to examine whether transparent communication and explanation of AI use in journalism can lead to more credibility. In contrast to previous findings, we could show that explanations and transparency do not have an impact on the credibility. However, we found that the credibility of media companies and users' experience with social media as well as the AI experience positively impacts the trust in AI-generated content.

Keywords: artificial intelligence, journalism, credibility, AI-generated content, automated journalism, trust, transparency.

1 Introduction

Social media are nowadays often used as a source of news and information. However, despite the high level of distribution and use of social media, its credibility is relatively low (Neuberger, Nuernbergk and Rischke, 2009). According to the credibility paradox, especially young people, on the one hand, increasingly use social media to search for information about current events (Stieglitz *et al.*, 2019; Mirbabaie, Bunker, *et al.*, 2020). On the other hand, they also adopt a critical attitude towards news published and distributed on social media (Preuß *et al.*, 2017). One reason for the low credibility and trustworthiness of social media is the high prevalence of fake news which has been on the rise since 2016, for example in political debates (Grimm, Keber and Zöllner, 2017). In some cases, these news which can be created and published by anyone are indistinguishable from factually correct news (Gelfert, 2018). In the past, fake news were often distributed through social bots (Ciampaglia *et al.*, 2018). Social bots are computer programs that use an algorithm to mimic human behavior by creating, commenting on and sharing social media posts in an attempt to steer users' opinions in a particular direction (Kind *et al.*, 2017; Brachten *et al.*, 2018). However, the use of algorithms in media is not limited to distributing fake news. Media companies also use artificial-intelligence-based systems (AI-based systems) for the autonomous creation of news content (Galily, 2018; Jones and Jones, 2019). In most cases, artificial intelligence' (AI) use in journalism is limited to the creation of texts with a low degree of complexity and high degree of standardization such as sports and financial reports (Graefe, 2016). However, the complexity of texts that can be generated by AI-based systems is increasing (Blankespoor, DeHaan and Zhu, 2018; Brennen, Howard and Nielsen, 2020).

As media companies have an impact on the opinion forming process of societies (Plaisance, Skewes and Hanitzsch, 2012) and their business model relies on being perceived as a credible source (Haim and Graefe, 2017), it is important that people trust the content published by journalists. However, research on the perception of AI-generated news content is still not fully examined. Research proposes transparency rules that show how an algorithm weights information, evaluates and allocates the data sources used (Diakopoulos and Koliska, 2017; Jones and Jones, 2019). Likewise, the origin and mode of operation of how an algorithm was trained should be disclosed (Diakopoulos and Koliska, 2017). Furthermore, a labeling requirement is called for, which makes both the identity and the origin of a non-human author recognizable (Mittelstadt, 2016). There is also a clear tendency towards transparent communication on how to deal with social bots and fake news on social media, i.e. when news is created by a human and when by a computer program (Preuß *et al.*, 2017). However, AI is not one specific technology, rather it is a group of technologies that can be applied in different contexts (Thurman, Lewis and Kunert, 2019). This raises the question of how different content that is created by an AI-based system is perceived by the public in terms of its credibility. There is also the question of how the credibility of journalistic articles in social networks differs from the credibility of articles on official news websites. Credibility is the key to success in journalism and AI will be increasingly used in media companies. Therefore, we pose the following research question: *What influence does the transparent communication of computer-generated news articles have on the credibility of the content displayed on social media sites and websites?*

To address this question, we conducted an online survey (n = 122) on the perception and evaluation of the credibility of AI-generated news articles. We followed the layer model of Lucassen and Schraagen (2012) which state that the credibility of information depends on the credibility of the source and the credibility of the medium. We presented news articles of different domains that were highlighted as created by an AI-based system or as written by a human journalist. We applied a 2x2 design and presented, on the one hand, articles on social media sites and on official news websites and, on the other hand, provided a detailed explanation of how the AI-based system generated the article or left this out. Up to now, the perception and evaluation of computer-generated news reports have been presented exclusively in a way that is not linked to the source and its medium. This work aims to investigate the assumption that the source is an important factor in the credibility assessment of AI-generated news content. We provide insights into the perception of AI-generated news content in relation to the source and the medium. We show how credibility can be measured in terms of multiple levels and provide knowledge on the targeted use of AI in media for information systems (IS) research as well as for practitioners.

2 Background

2.1 Automated Journalism

Automated journalism is not only characterized by automated information processing and publishing, but also by the fact that computer programs can create journalistic articles independently (van Dalen, 2012; Lokot and Diakopoulos, 2016). The central element of automated journalism is an algorithm that researches data from current news databases, evaluates and analyzes them, and generates independent news articles from the data using pre-programmed text modules (Graefe, 2016). Automated journalism is often also called 'algorithmic journalism' or 'machine-written journalism' (Thurman, Doerr and Kunert, 2017; Lewis, Guzman and Schmidt, 2019). This avoids the misleading term 'robot journalism', as it is not a robot but a computer program that has to be programmed, developed and maintained by a human (Blankespoor, DeHaan and Zhu, 2018). Until now, automated journalism has specialized in the creation of short textual news and is only used by a few media companies such as Associated Press from the US (Graefe *et al.*, 2018). Applications include automated traffic and weather reports as well as financial and sports reports. These are text formats that are fact-based and have a more objective language style (Tatalovic, 2018).

The automated journalism described in this work is based on an AI-based system. This *"refers to the ability of a machine to perform cognitive tasks that are linked to the human mind. This includes possibilities for perception, as well as the ability to reason, to learn independently and thus to find solutions to problems independently"* (Kreutzer and Sirrenberg, 2019, p. 3).

In journalism, AI-based systems collect data on current events in a database and classify the most important and interesting information using various tools and methods (Graefe *et al.*, 2018). The classified data is then transferred to a natural language generation system which analyzes and interprets the data according to predefined rules. For the general public, it becomes increasingly nontransparent how articles were created. However, the explainability of AI is of high relevance for achieving a higher credibility (Kim, Park and Suh, 2020; Sachan *et al.*, 2020).

2.2 Perception of AI-generated content

How users perceive computer-generated texts has already been investigated in a number of studies from different countries (Clerwall, 2014; Mirbabaie, Stieglitz, *et al.*, 2020). Clerwall (2014) showed in one study that Swedish students were not able to clearly identify the author of the text, nor did they find any differences in the assessment of AI-generated content in comparison with human-written texts in terms of credibility and readability. However, one explanation could be that both the presented computer-generated texts and the texts written by humans were found to be not interesting and not very pleasant to read (Clerwall, 2014). This indicates that considering only the author does not provide a clear conclusion about the perceived credibility of an article. In general, it seems to be not relevant for the evaluation of articles whether a human or a computer was indicated as author of a text (van der Kaa and Krahmer, 2014). Also in specific contexts such as soccer news, texts that were declared to have been written by a journalist were rated only slightly higher in the categories credibility, readability and journalistic expertise than the texts whose authorship was declared to be computer-generated (Graefe, 2016). The actual author, i.e. whether the text was written by a human being or computer-generated, was irrelevant. If, however, only the actual source is considered, minimal contrary effects were observed in the categories credibility and journalistic expertise, because regardless of the topic and the stated authorship, the computer-generated texts were recognized as having a higher credibility and more journalistic expertise (Graefe, 2016). Expectations could also not be found as an indicator for the subsequent evaluation (Haim and Graefe, 2017).

It could also be assumed that subjects who are more familiar with the topic presented would have higher expectations regarding factuality and objectivity, both of which are the subject of the items of credibility. However, Graefe et al. (Graefe *et al.*, 2018) found that the topic involvement has no moderating effect on the perception of news. Haim and Graefe did not completely reject the topic involvement as a moderating effect and call for further research on how highly involved recipients perceive computer-generated texts that are tailored to their personal needs (Graefe *et al.*, 2018).

Not all previous studies confirmed that AI-generated articles are perceived as credible as human-written articles. Wölker and Powell (2018) found that sports articles were rated higher in terms of credibility and readability of the computer-generated texts. They suggest that further research should therefore compare different topics and areas such as celebrity and political news (Wölker and Powell, 2018). Overall, the previous studies show some weaknesses or incompleteness, as they measured either articles with a specific topic or only parts of the credibility.

3 Credibility, Transparency and Trust in Journalism

Credibility in journalism has declined, especially with the establishment of social media (Kunert, Hofrichter and SimonAnja, 2019). One key to more trust in journalism is transparency of how the article was created and of who created the article (Kovach and Rosenstiel, 2007). Transparency is not only ethically desirable (Diakopoulos and Koliska, 2016; Mittelstadt, 2016; Brendel *et al.*, 2021), it can also increase credibility in journalistic articles (Meier and Reimer, 2018). Meier and Reimer (2011) showed that in the case of print articles, product transparency leads to greater trust, whereas process transparency leads to greater trust in online articles. As a result, the editorial openness should be emphasized, especially in online journalism (Meier and Reimer, 2018).

Another credibility assessment of journalistic contributions in online environments can be explained by heuristic processes. Not every piece of information is checked in detail as this would require a lot of effort. As a solution, certain clues are being applied from which credibility can be inferred (Taraborelli, 2008; Metzger, Flanagin and Medders, 2010). Lucassen and Schraagen (2011) examined which clues could be applied for credibility assessment. In their 3S-model they showed that the most direct strategy to assess credibility is the search for semantic clues in the information itself. This takes into account indications such as factual accuracy, neutrality or completeness of the information. In addition, the source or the medium itself which is used to disseminate the information can also be evaluated heuristically (Lucassen and Schraagen, 2012). As an extension of their 3S-model Lucassen and Schraagen (2012) derived a layered model in which they distinguished between trust that is built from a general tendency to trust to a case-specific trust in a particular piece of information. In this model, the general tendency to trust is regarded as the general baseline of a person's trust in all situations, not only for trust in online environments. The second level is called trust in the medium. It is a generalization and describes the trust of a recipient in a certain type of media, such as newspapers, radio or the internet in general. The next level describes the trust in the information source and only then follows the trust in the information itself. Only if the credibility of the source is doubtful do the recipients look for clues in the information itself to assess its credibility (Lucassen & Schraagen, 2012). In addition to Lucassen and Schraagen (2012), further studies have shown that the credibility assessment depends not only on personal factors, such as the general tendency to trust (Johnson and Kaye, 2002), but also on the frequency of use of a medium. People rate media they prefer as a source of information more often as more credible than media they use less frequently (Mehrabi, Hassan and Ali, 2009). Furthermore, credible websites are characterized by reporting on current events and transparently communicating who wrote the respective article (Hong, 2005).

Although the source of the information, the subject and the way the information is presented are important variables in the context of credibility, social aspects may also have a great influence in judging the credibility of an article created by an AI or a journalist. As one example, gatekeeping theories (Singer, 2006, 2014; Shoemaker and Vos, 2009) and network gatekeeping theories (Barzilai-Nahon, 2008; Ernste, 2014; Deluliis, 2015) have been a popular heuristic for describing information control (Barzilai-Nahon, 2008). Gatekeeping in journalism can be described as a process of controlling information through a filter by journalists or editors (Barzilai-Nahon, 2009). As gatekeeping theories focus on the shifting role of journalists and we aim to investigate the perception of AI-journalism by the general public, we were inspired by the layered model of Lucassen and Schraagen (2012) as a comprehensive basis to examine trust in AI-generated news articles. We consider gatekeeping as one influencing factor in the context of the credibility of the source.We also challenge the model of Lucassen and Schraagen (2012) in some parts because it highly simplifies the formation of credibility. With respect to AI, it cannot be assumed that the process of perceiving AI-generated content is linear layer by layer. In our study, therefore, the information and the author are shown simultaneously. We argue that the structure of the model is not crucial and that context factors are more important for credibility such as the frequency of use of a medium or the experience with AI. Trust is not a linear process in which news recipients only proceed to the next level if they have established trust in the previous level.

4 Derivation of the Hypotheses

In order to be able to process the increasing amount of information, media companies will increasingly use AI-based systems to support them in writing articles. Thereby it must be ensured that readers have confidence in such content. The literature showed that recipients do not recognize any difference between computer-generated and human-written articles in most cases (Clerwall, 2014; Graefe *et al.*, 2018). Neither the actual source nor the marked authorship has a major influence on the perception and evaluation with regard to the credibility, readability and journalistic expertise of the contributions (Clerwall, 2014; van der Kaa and Krahmer, 2014; Haim and Graefe, 2017; Jung *et al.*, 2017). Similarly, expectations and the topic involvement have no influence on this perception of the contributions (Graefe *et al.*, 2018; Haim and Graefe, 2018). However, the articles written and published by algorithms and thus also used in the studies so far are also very simply written news items that always followed the same scheme, as for example in the case of financial and sports news. Considering that these news items consist of a simple recitation of facts and that the subjects often lack a sophisticated narrative, it is not surprising that the recipients rated the articles as quite credible and knowledgeable. However, the subjects did not like reading both types of articles. One explanation for the low readability rating could be that sports and finance are specific topics that are not interesting for everyone (Graefe, 2016). Haim and Graefe point out that there is a need for research into the selection of topics in automated journalism. They also highlighted that especially in online journalism, more transparency is demanded regarding the creation of articles, when and how an algorithm can be used. Based on this demand for transparent communication, the first hypothesis can be derived: *H1: The transparent communication of the author has a positive influence on credibility of AI-generated news articles.*

Likewise, the contributions in the preceding studies were shown uncoupled from a source. This means that the contributions were presented independently of a medium such as a website, app or social media and independently of a specific news provider. However, since internet news websites and especially social media are becoming increasingly important as a source of information, it is highly valuable to investigate how the credibility of the respective websites of news providers compared to the respective social media channels affects the credibility of news articles created by an AI-based system. According to the layer model of Lucassen & Schraagen (2012), it can be assumed that the credibility of the source has an influence on the credibility of the news information which leads to our second hypothesis: *H2: The credibility of the source (e.g. a certain news provider) has a positive influence on the credibility of the presented AI-generated news article.*

Furthermore, in their layer model Lucassen and Schraagen (2012) concluded that the credibility of an information is based on the trust of a recipient in a certain type of media, social media and websites. Only if a user trusts the medium the source is trusted. Social media in particular is considered as less credible than official websites of news providers (Ernste, 2014; Ciampaglia *et al.*, 2018; Newman, 2018). One explanation is that the necessary credibility and trustworthiness of social media is reduced, especially with regard to the distribution of fake news and misinformation. Misinformation and false information, which is deliberately generated and especially published online, manipulate the public, with social bots often serving as a tool (Fairfield and Shtein, 2014). We therefore assume that the credibility of information presented on social media differs from the credibility of information on news websites: *H3: The credibility of information (the article's content) is evaluated differently on news websites than on social media due to transparent communication.*

Especially non-users of social media are very skeptical and do not see any connection, or only a small one, with journalistic quality (Neuberger, Langenohl and Nuernbergk, 2014). As one possible explanation, social media users have a clear idea of the role of journalism on social media sites. Due to their experience for them it is easier to recognize the differences. Furthermore, it could be shown that the frequency of use of news websites and social media channels influences the credibility of news articles (Neuberger, Langenohl and Nuernbergk, 2014), which led to a fourth hypothesis: *H4: The frequency of use of news websites and social media channels correlates positively with the credibility of AI-generated news articles.*

With these four hypotheses we aimed to examine the influence on trust in AI-generated news articles in order to provide guidance for the use of AI-based systems in practice and to contribute to the research stream of AI trust in the domain of journalism.

5 Method

In order to examine the impact of highlighting articles as AI-generated on trust in social media and websites, we conducted an online survey. As target group, we recruited 122 internet and social media experienced and not-experienced participants of different age and educational background through social media channels such as Facebook, Instagram, and LinkedIn. In a preliminary study, the credibility of eight German news providers (BILD, Tagesschau, RTL, FAZ, WDR, ZDF, Stern and Spiegel Online) was tested with a 5-Point-Likert scale. In addition, 24 headlines of emerging topics were presented. According to the prestudy, we selected 13 topics as stimulus materials for the main study.

For the main study, we followed Haim and Graefe (2018) and used a 2x2 between design to examine the difference between transparent communication of using AI to create content on social media and news websites. The medium varied by presenting the journalistic contributions either as screenshot on the social media application Instagram or on an official website of a news provider. We chose Instagram as a representative social media platform because the platform has seen a high increase in usage in recent years and media companies are highly active on the platform (Vázquez-Herrero, Direito-Rebollal and López-García, 2019). Although Instagram is actually a platform for images, it is increasingly being used by media providers as their primary social medium to distribute short news articles. Younger news recipients in particular are more likely to receive their news through Instagram than through other social media platforms (Vázquez-Herrero, Direito-Rebollal and López-García, 2019). Since news on websites also tend to be short, we decided to compare texts of news websites with Instagram news. According to Haim and Graefe (2018), we did not present real computer generated articles, but just declared certain articles as AI-generated. Articles generated by a sophisticated AI should be indistinguishable from human articles, thus it was not relevant to the aim of this study whether the articles were really generated by AI. As we could not guarantee that real computer-generated texts would not slightly differ from the human texts in terms of content, we decided to declare the same texts for one group of participants AI-generated in order to endure content level comparability.

We also considered whether the frequency of use of social media has an influence on credibility. Since all eight news providers tested in the pretest were regularly used to obtain information about current events, we followed the layer model of Lucassen and Schraagen (2012) to examine whether the credibility of the source has an influence on the rating of computer-generated texts (Table 1)

Table 1. Representation of layers from Lucassen and Schraagen (2012) in the study

Layers	Representation of layer	Questionnaires
Trust in the information	Short news articles tested for relevance in the pre-study	(Sundar, 1999)
Trust in the source	Diverse news providers' logos	Self developed questionnaire in pre-test
Trust in the medium	Screenshots from news providers' website or Instagram page	(Lucassen and Schraagen 2014)
Propensity to trust	Different propensities were identified by questionnaire	(Costa, P T. & McCrae, 1992)

As stimulus materials, we selected thirteen journalistic articles evaluated in the prestudy. Each of these articles were published on Instagram and the news provider's website. The articles were about six weeks old at the time of publication to ensure that they were related to the current topic. Unlike the previous studies, however, it was clear from which news provider they were published and whether they were published on the website or on Instagram.

In order to investigate the difference between transparent communication and the use of AI, four randomized groups were formed which differed both in the medium and in the transparent communication (Table 2). The transparent groups 1 and 3 were first provided a definition and a figure of an AI-based system in journalism. Afterwards we presented a sample text (see Figure 2) which reported about the first soccer game without fans in Germany. This text was provided with the exact information which passages were AI-generated. The provided information is presented in Figure 1.

AI searches for data and saves them in a database

AI configures, organizes, evaluates data and concludes from them

AI writes content with pre-programmed text modules

We now present a short exemplary article written by an AI-journalist. The bold passages were taken from a database and adapted to the pre-programmed text modules.

*After a **2-0 lead, the Rhinelander** missed the victory in the first ghost home game in history of the club against **FSV Mainz 05** and missed the chance to get close to the European Cup spots. Although, the class retention should be fixed at **ten points ahead of relegation rank 16**, **six points** separate the **FC** from the **coveted sixth place**. Meanwhile, the **Mainz** team is still **four points behind 16th place**.*

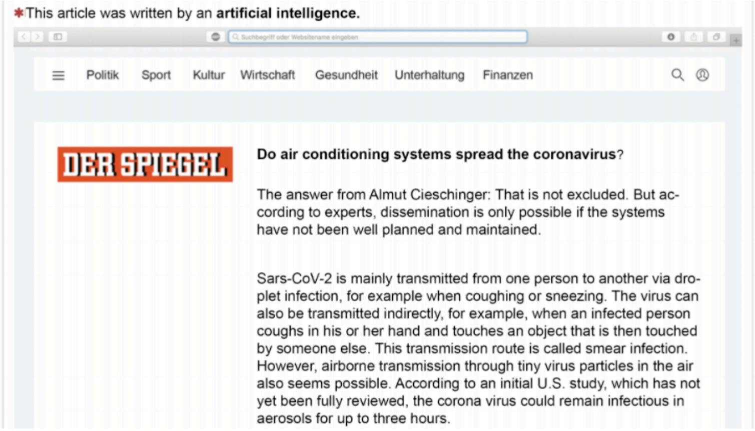
Figure 1. The illustration and exemplary text presented to the participants from groups 1 and 3

The participants either received instructions on how an AI writes an article and then perceived the articles that were published either on Instagram or on the website, or they were able to view the articles directly without any instructions. To illustrate the differences, all groups saw the same articles. These were either erroneously declared to have been written by an AI, or declared to be written by a human journalist (as it was the case). These authorships varied within the groups so that participants were presented both articles written by an AI-based system and by human authors. At the same time, we reminded the participants to pay attention to the authorship when reading the article.

Table 2. Randomized groups in the online survey that were presented AI / human written articles

	Platform	**Author transparency through definition, example and figure**
Group 1	Instagram	Yes
Group 2	Instagram	No
Group 3	Website of news provider	Yes
Group 4	Website of news provider	No

The participants were asked to rate the credibility, readability and journalistic expertise on a five-point-scale using the items according to Sundar (Sundar, 1999).



*This article was written by an **artificial intelligence.**

Do air conditioning systems spread the coronavirus?

The answer from Almut Cieschinger: That is not excluded. But according to experts, dissemination is only possible if the systems have not been well planned and maintained.

Sars-CoV-2 is mainly transmitted from one person to another via droplet infection, for example when coughing or sneezing. The virus can also be transmitted indirectly, for example, when an infected person coughs in his or her hand and touches an object that is then touched by someone else. This transmission route is called smear infection. However, airborne transmission through tiny virus particles in the air also seems possible. According to an initial U.S. study, which has not yet been fully reviewed, the corona virus could remain infectious in aerosols for up to three hours.

Figure 2. Exemplary screenshot of a news item on a news provider's website presented in the study

Prior to this, the frequency of use of news websites, social media and Instagram was surveyed in accordance with Preuß et al. (2019). In addition, we examined the trust and frequency of use of websites and social media sites based on the scale of Lucassen and Schraagen (2014). The scale was adjusted to a five-point-Likert scale. The items "use", "perceived credibility", "trust in institutions and individuals", "usefulness" and "privacy" were retained. Furthermore, we added the questions of the net-confidence and net-risks scale by Dutton & Shepherd (2006). We also added questions of whether news websites are more likely preferred and whether the risk of receiving false information from news websites is perceived (Kari Kelton, Kenneth R. Fleischmann, 2007). We surveyed the trust and frequency of use of social media with the same five questions.

Cronbach's alpha for trust in Internet sites was $\alpha=.70$, which indicates good reliability. For social media trust, Cronbach's alpha was α=.824, indicating a very good reliability. The propensity to trust was measured using the NEO-PI-R personality test by Costa and McCrae (1992). As some personality traits are not relevant for this survey, we did not use all of them in our study. The eight items were operationalized on a five-point-Likert scale. Although the NEO-PI-R questionnaire is not intended for partial use, Cronbach's alpha with $\alpha=.831$ showed very good reliability for the remaining questions. Items that were considered to be suspicious, cynical and skeptical were reversed coded.

We conducted the participants' experience with AI on six independently selected items on a five-point Likert scale. We conducted the level of knowledge about what an AI-based system is and the experience as well as the well-being when interacting with an AI. In addition, we measured the personal attitude and critical opinion whether AI should be used in journalism with three items. Since the items represent the basic attitude which will be discussed later in this paper we did not include it in further calculations. For the other six items, Cronbach's alpha was $\alpha=.786$ and thus shows a very good reliability. In order to ensure that the authorship was perceived, we checked whether the participants were aware that some articles were written by an AI-based system. After that, the participants were presented a questionnaire on the general tendency to trust. Finally, we conducted demographic data on gender, age, education and occupation.

A total of 122 persons participated in the study who were selected randomly and without exclusion criteria. Data sets for which participants answered "no" to the control question whether they were aware that some texts were written by an AI-based system were excluded afterwards. As a result, the groups were unbalanced and had to be adjusted. The final sample of $n = 84$ participants was divided into 44 women and 40 men. The average age of the participants was 36.57 years.

6 Findings

The questionnaires, including the evaluation of the contributions, were collected with LimeSurvey. The data was first cleaned up in Excel and the items were recoded, which were formulated in the opposite way. These were two items in the NEO-PI-R personality test. Subsequently, the items for credibility, readability and journalistic expertise were combined by calculating the overall mean value for each respondent. The same applies to the individual news providers. Here it was important to note that not only the number of articles was different for each news provider, but also that authorship varied in the different groups. We a priori determined a significance level of *0.05*. For each of the variables collected, the most important statistical parameters were calculated and then tested for normal distribution.

Only the variables of the confidence questionnaire for social media $(p = .154)$ showed no significance under Shapiro-Wilk with a significance level of $p \leq .05$, so that a normal distribution can be assumed here. The variables of the confidence questionnaire for internet media $(p = .021)$, as well as those of the NEO-PI-R $(p = .041)$ and those for the experience with AI $(p = .009)$ did not show a normal distribution under Shapiro-Wilk (Shapiro and Wilk, 1965). Nevertheless, almost all points are on the straight line of the Q-Q plot both for the variables of the confidence questionnaire for internet sites and for the NEO-PI-R personality test and the experience questionnaire. Moreover, the sample consists of more than 30 subjects, which is why a normal distribution can be assumed under these conditions (Chambers *et al.*, 1983; Fowlkes, 1987). The results showed that while transparent communication did not have a significant difference in the assessment of credibility, a tendency towards higher scores can be seen when the use of AI in journalism is communicated *($t(82) = 1.75$, $p = .084$)*. A mean value comparison for the credibility rating on news websites and social media did not show any statistically significant difference with regard to transparent communication *($t(40) = 1.106$, $p = .276$, $t(40) = 1.335$, $p = .189$)*. In both tests, however, the contributions were rated 0.265 units better on average in the transparent group (Table 3).

Table 3. Results of t-tests

	t	**df**	**P (two-sided)**
Trust in AI-generated articles (general)	1.751	80	.084
Trust in AI-generated articles (websites)	1.106	40	.276
Trust in AI-generated articles (Instagram)	1.335	40	.189

However, we found an influence of the source and frequency of use of social media on the credibility rating. The overall comparison of the sources with each other showed a highly significant result with a significance level of *$p \leq .05$, $F(5.10,423.66) = 16.20$, $p < .001$, $\eta^2 = .16$.* According to Cohen (1988), with a value of *$f = 0.44$*, a medium effect size could be found. The descriptive statistics also showed that the credibility of the sources Tagesschau, ZDF and FAZ was rated significantly better than the sources BILD and RTL. Thus, the articles were rated better by a credible source in terms of credibility than by an untrustworthy source. The results from a Bonferroni post-hoc test are shown in Table4.

Table 4. Bonferroni post-hoc test for comparing trustworthiness

Source 1	**Source 2**	**Av. difference**	**Standard error**	**P value**
BILD	Tagesschau	-.470	.81	.000
BILD	FAZ	-.531	.113	.002
BILD	ZDF	-.574	.072	.000
RTL	Tagesschau	-.753	.100	.000
RTL	FAZ	-.649	.100	.000
RTL	ZDF	-.530	.077	.000

To analyze whether the frequency of use of Instagram, social media and of news websites has an influence on the assessment of the credibility of AI-generated news articles, an ordinal regression was performed, since neither the use of Instagram and other social media, nor the use of news websites had a linear relationship and (with a significance level of *$p \leq .05$*) a significance under Shapiro-Wilk was found. The results of the ordinal regression analysis on the frequency of use showed that the use of Instagram *(Chi-square (1) = 7,353, $p = .007$, $n = 84$* and Social Media *Chi-square (1) = 8,734, $p = .003$, $n = 84$)* had an impact on the credibility rating, whereas the use of news websites had no impact on the rating *(Chi-square (1) = . 789, $p = .374$, $n = 84$)*. The position estimators of ordinal regression also show significance for the variables of Instagram usage frequency with respect to the evaluation of credibility. Thus, the more frequently social media is used, the more credibly the contributions are rated (Table 5).

Table 5. Results of ordinal regression of trust in AI-generated articles

	Estimate	**Standard Error**	**Wald**	**df**	**Sig.**
Website use	-.228	.245	.862	1	.353
Social media use	.476	.156	9.239	1	.002
Instagram use	.293	.107	7.422	1	.006

Thus *H2* and *H4* were confirmed, whereas *H1* and *H3* showed no significance but a tendency.

7 Discussion

The results of this study showed that there was no significant difference if the perticipants were informed beforehand how AI-generated articles were created (*H1* not significant). Credibility, readability, and journalistic expertise, independent of transparent enlightenment were similarly evaluated. However, the groups that received an explanation of how an AI works scored slightly better in all three variables, regardless of whether the articles were presented on social media or on the internet. This is in strong contrast to explainable AI (Kim, Park and Suh, 2020; Sachan *et al.*, 2020), which aims to create transparency and trust by explaining how AI works, as in our study we did not find significant differences. We explain this by the fact that articles generated by AI are already considered quite credible. As a result, highlighting how an AI-based system created an article just slightly increases credibility. In fact, following on from previous research (Mittelstadt, 2016; Haim and Graefe, 2017; Preuß *et al.*, 2017), we found that articles written by human journalists were only minimally more credible, readable, and expertly rated than articles generated by an AI-based system. The actual content seems to be much more important than the authorship when assessing the credibility of an article. In addition, social theories such as gatekeeping theories (Barzilai-Nahon, 2008; Shoemaker and Vos, 2009; Singer, 2014) and network gatekeeping theories (Ernste, 2014; Deluliis, 2015) may have a greater influence in judging the credibility of news articles created by an AI or a journalist than the authors themselves. Possibly, the content of the news cannot be considered separately from the author. Lucassen and Schraagen (2012) assume different layers in their model. Our results suggest that these layers should not be considered separately and that additional gatekeeping theories should be considered as social theories to understand credibility in this context.

Interestingly, the content presented on social media were rated better by the transparent group in all areas than the presentation on the respective website of news providers. With regard to the assumption that social media in particular are classified as less credible due to the high prevalence of fake news and misinformation (Ciampaglia *et al.*, 2018), we could show that process transparency can lead to more credibility. This assumption could also be confirmed by the evaluation of the non-transparent group which rated the articles presented on the websites better than those published on social media. Perhaps this is also related to the changing role of journalists as gatekeepers (Deluliis, 2015), where users gain much more control over content as they can create content themselves and become secondary gatekeepers by upgrading or downgrading posts through their engagement (Singer, 2014). However, we interpreted these results with caution as they have not become significant (*H3*). Nevertheless, it can be deduced for journalism that a clear process transparency of how an article is created could lead to more credibility on social media as a source but not to more credibility of the author.

Furthermore, we found a highly significant correlation between the credibility of a source, i.e. the credibility of a news provider, and the credibility rating of AI-generated articles (*H2* significant). For example, the articles which were declared to be written by an AI-based system were rated better for credible news providers such as Tagesschau, ZDF or FAZ than for non-credible providers such as BILD or RTL. As a conclusion, providing an explanation of how AI creates articles is of high relevance for credible media companies whereas it is less important for tabloid press. This could be due to the fact that with more credible media providers, readers are more likely to expect them to provide explanations, while readers of tabloid media are less likely to expect this. It could also result from more in-depth reports from credible media providers, where readers are more likely to seek explanations for their origins.

With regard to the model of Lucassen & Schraagen (2012), we did not examine the information from a non-credible source for any further indications to conclude on credibility. As an example, the articles of the non-credible news providers were nevertheless rated with a medium high credibility. This can be explained by the topic of the presented articles. In contrast to the previous studies (Haim and Graefe, 2017; Graefe *et al.*, 2018), articles on topics from the fields of politics, business and entertainment were selected in this study. It could be observed that articles from the fields of politics and economics were generally better rated than the articles from the field of entertainment, regardless of the medium and source. In order to reinforce the model's statement, we nevertheless found that both articles from the field of economics published by BILD and RTL were rated worst of all articles in terms of credibility. This means that although the credibility of the source is very low, the information is examined for further clues to conclude its credibility. According to the layer model of Lucassen and Schraagen (2012), not only the source has an important influence on credibility, but also the subject matter and form of presentation of the information itself. Lucassen and Schraagen (2012) also assume that trust in a medium also has an influence on the credibility of the source and thus on the credibility of the information itself. However, they found no significant correlation between trust in social media and the assessment of credibility. We found the same for trust in news websites.

Overall, we were able to show that the credibility of news articles cannot be represented by a linear processing of the layers of Lucassen and Schraagen (2012). Contextual factors such as experience with AI and the frequency of use of social media play a far more important role. The frequency of use of social media, and therefore also of Instagram, has a significant influence on the assessment of credibility (*H4*) whereas the frequency of use of news websites shows no significance. This was very surprising, because previous studies indicated that social media is increasingly being used as the main news source, especially by young adults, and Instagram in particular is gaining in importance (Kunert, Hofrichter and SimonAnja, 2019). However, it is precisely those under 30 years of age who have a high level of usage who rated the articles as less credible. Additionally, we could show that especially those under 30 years of age who have more experience with AI-based systems were more critical. We assume that young adults in particular who are more likely to use social media and thus come into contact with AI in the form of social bots more often, are more aware of the negative consequences. In addition, their knowledge of what an AI can do might influence the assessment, because according to the current state of knowledge, an AI is not able to independently develop sensory connections, to adequately grasp topics or even to develop its own point of view (Graefe *et al.*, 2018).

Meier and Reimer (2018) particularly demanded process transparency in online journalism in order to achieve more credibility. This demand came together with the proposal of Mittelstadt (2016) to introduce a labeling requirement for the use of an AI that discloses the procedure of the algorithm. In addition, there should be an opportunity to contact the editors if the users have questions or criticism. Even if these aspects do not lead to more credibility of the author, they are still highly relevant from a media-ethical point of view. Online media as well as news providers in general play an important role on the opinion formation process of people and they bear a social responsibility for societies. That is why the criteria of transparency, traceability, non-discrimination and verifiability should be clearly recognizable when an AI is used (Grimm et al., 2017). However, as Sieber (2019) already mentioned, especially the traceability of an algorithm is not always obvious. Data quality is not evident when using AI (i.e. which databases are searched by AI for reporting purposes). Therefore, in order to achieve greater credibility, journalists should rely on contextual factors of trust and distinguish between different target groups (e.g young social media users and older website users), as we could not find any influence of transparent communication of how AI works in journalism on the credibility of articles in this study.

8 Limitations and Future Research

This study comes with some limitations. We had to exclude 17 percent of the participants as they were not aware that some articles were written by an AI-based system. This suggests that the indication of whether the article was written by a journalist or an AI-based system could be even clearer. Although there was an adequate division among respondents by gender, we did not focus on the results associated with male or female gender even though this could be another contextual factor that influenced perception. Future studies may look more closely at this relationship. It might also be interesting to divide participants into further age groups (e.g., generations X, Y, and Z). By showing that the credibility of the authors was higher on social media than on news websites, future research could also consider the role of users as secondary gatekeepers (Singer, 2014) as an extension of Lucassen and Schraagen's (2012) model.

We chose Instagram as a representative social media platform as we found a high increase in usage in recent years and media companies are highly active on the platform (Vázquez-Herrero, Direito-Rebollal and López-García, 2019). However, older generations still make intensive use of other social media platforms. Future studies should consider other social media platforms such as Facebook or YouTube and could observe the credibility of information between those platforms in a comparative study. Furthermore, it would have been interesting to compare a contribution that was actually written by an AI, as shown in the previous study, with the falsely declared texts. This would also have revealed the areas of sports and finance in comparison to politics, economics and entertainment. Also interesting would be the dimension of trust in the correctness of the presentation of information. This also examines credibility in journalism by asking whether the article would stand up to scrutiny and whether it presents the facts as they are. Finally, future research should expand this research to other cultures and countries as it is likely that there are differences (with e.g., underdeveloped countries).

For the IS community, this means that further research should be conducted, especially in this area to explore why the transparent communication of AI's process of content creation in journalism does not impact its credibility. Especially how the use of AI can lead to more trust and credibility among users should therefore be explored.

9 Conclusion

In relation to our research question, it can be summarized that a transparent communication about the use of AI-generated news content displayed on social media sides and news websites does not show a significant influence on the credibility of the content. This may be due to transparent communication by the author not being sufficient to generate credibility, but rather an interplay of various contextual and social factors being crucial.

Overall, with this study we provide some contributions to the research and practice. Although we found a slight tendency, news providers in general cannot increase their credibility through communicating how an applied AI-based system creates an article. This implies that media organizations cannot only rely on transparency to increase their credibility. Rather, contextual and social factors such as the changing role of journalists as gatekeepers and users as secondary gatekeepers need to be considered in more detail. This finding stands in contrast to the explainable AI literature and is an important contribution to IS research. However, we could show that the credibility of the source positively impacts the trust in AI-generated content. Therefore, one can conclude that using AI-based systems in journalism is more suitable for credible media companies than for tabloid press. As a further contribution to research, we showed that the source has a great influence on the evaluation of journalistic articles as this is often heuristic which is also confirmed in this study. Furthermore, the more often social media is used, the better AI-generated articles are rated. This could be associated with the fact that users who use social media frequently have a certain expectation of journalism on social media and thus also place greater trust and credibility in it. This can also mean that the transparent communication of AI-generated content in journalism will be more important in the future as more people will be experienced with using social media.

On social media, the credibility of articles was better assessed if it was clearly communicated that and how an AI generated the texts which show the potential for AI use on platforms such as Instagram. This may be due to the shifting role of journalism as gatekeepers and users as content creator and secondary gatekeeper, but will rather be rooted in the fact that we should not consider individual levels of Lucassen and Schraagen's model separately, but in interaction and in connection with social and contextual factors such as users' previous experience with AI and the frequency of use of social media. Transparency is one of the most important quality characteristics in journalism and for the use of AI-journalists it was assumed to be very important. However, it is not a panacea. We could show that providing explanation of how AI creates articles in journalism does not significantly impact the credibility of the content.

References

Barzilai-Nahon, K. (2008) 'Toward a Theory of Network Gatekeeping: A Framework for Exploring Information Control', *Journal of the American Society for Information Science and Technology*, 59(9), pp. 1493–1512. doi: 10.1002/asi.

Barzilai-Nahon, K. (2009) 'Gatekeeping: A Critical Review', *Journal of Chemical Information and Modeling*, 43, pp. 1–79.

Blankespoor, E., DeHaan, E. and Zhu, C. (2018) 'Capital market effects of media synthesis and dissemination: evidence from robo-journalism', *Review of Accounting Studies*, 23(1), pp. 1–36. doi: 10.1007/s11142-017-9422-2.

Brachten, F. *et al.* (2018) 'Threat or opportunity? - examining social bots in social media crisis communication', *arXiv*.

Brendel, A. B. *et al.* (2021) 'Ethical management of artificial intelligence', *Sustainability (Switzerland)*, 13(4), pp. 1–18. doi: 10.3390/su13041974.

Brennen, J. S., Howard, P. N. and Nielsen, R. K. (2020) 'What to expect when you're expecting robots: Futures, expectations, and pseudo-artificial general intelligence in UK news', *Journalism*. doi: 10.1177/1464884920947535.

Chambers, J. M. *et al.* (1983) *Graphical methods for data analysis*. CRC Press.

Ciampaglia, G. L. *et al.* (2018) 'Research Challenges of Digital Misinformation: Toward a Trustworthy Web', *AI Magazine*, 39(1), pp. 65–74. doi: 10.1609/aimag.v39i1.2783.

Clerwall, C. (2014) 'Enter the Robot Journalist: Users' perceptions of automated content', *Journalism Practice*. Informa UK Limited, 8(5), pp. 519–531. doi: 10.1080/17512786.2014.883116.

Costa, P T. & McCrae, R. (1992) 'Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Model (NEO-FFI) Professional manual', *Odesa, FL; Psychological Assesment Center*.

Costa, P. (2018) 'Neo PI-R professional manual', (January 1992). Psychological Assessment Ressources. Odessa, FL.

van Dalen, A. (2012) 'The Real Algorithm Behind The Headlines', *Journalism Practice*, 6(5–6), pp. 648–658. doi: 10.1080/17512786.2012.667268.

DeIuliis, D. (2015) 'Gatekeeping theory from social fields to social networks', *Communication Research Trends*, 34(1), pp. 4–23. Available at: https://search.proquest.com/docview/1667629711/fulltextPDF/C3BCD3DA4ADA4218PQ/1?accountid=14648%0Ahttp://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=101594304&site=ehost-live.

Diakopoulos, N. and Koliska, M. (2016) 'Algorithmic Transparency In The News Media', *Digital Journalism*. doi: 10.1080/21670811.2016.1208053.

Diakopoulos, N. and Koliska, M. (2017) 'Algorithmic Transparency in the News Media', *Digital Journalism*, 5(7), pp. 809–828. doi: 10.1080/21670811.2016.1208053.

Ernste, T. (2014) 'The networked gatekeeping process for news in the 21st century', *2014 International Conference on Collaboration Technologies and Systems, CTS 2014*, (May 2014), pp. 11–18. doi: 10.1109/CTS.2014.6867536.

Fairfield, J. and Shtein, H. (2014) 'Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism', *Journal of Mass Media Ethics*, 29(1), pp. 38–51. doi: 10.1080/08900523.2014.863126.

Fowlkes (1987) *A Folio of Distributions - A Collection of Theoretical Quantile-quantile Plots*. CRC Press.

Galily, Y. (2018) 'Artificial intelligence and sports journalism: Is it a sweeping change?', *Technology in Society*, 54(March), pp. 47–51. doi: 10.1016/j.techsoc.2018.03.001.

Gelfert, A. (2018) 'Fake news: A definition', *Informal Logic*, 38(1), pp. 84–117. doi: 10.22329/il.v38i1.5068.

Graefe, A. (2016) 'Guide to Automated Journalism', *Tow Center for Digital Journalism Report*, (January), pp. 1–48. doi: 10.1002/ejoc.201200111.

Graefe, A. *et al.* (2018) 'Readers' perception of computer-generated news: Credibility, expertise, and readability', *Journalism*, 19(5), pp. 595–610. doi: 10.1177/1464884916641269.

Grimm, P., Keber, T. O. and Zöllner, O. (2017) *Digitale Ethik*.

Haim, M. and Graefe, A. (2017) ‘Automated News: Better than expected?’, *Digital Journalism*, 5(8), pp. 1044–1059. doi: 10.1080/21670811.2017.1345643.

Haim, M. and Graefe, A. (2018) ‘Automatisierter Journalismus’, *Journalismus im Internet*, pp. 139–160. doi: 10.1007/978-3-531-93284-2_5.

Hong, T. (2005) ‘The influence of structural and message features on Web site credibility’, *Journal of the American Society for Information Sci-ence and Techno*. doi: https://doi.org/10.1002/asi.20258.

Johnson, B. T. and Kaye, B. K. (2002) ‘Reliance Predict Online Credibility’, *Journalism & Mass Communication QuarterlyMass Communication Quarterly*, 79(3), pp. 619–642.

Jones, B. and Jones, R. (2019) ‘Public Service Chatbots: Automating Conversation with BBC News’, *Digital Journalism*. Routledge, 7(8), pp. 1032–1053. doi: 10.1080/21670811.2019.1609371.

Jung, J. *et al.* (2017) ‘Intrusion of software robots into journalism: The public’s and journalists’ perceptions of news written by algorithms and human journalists’, *Computers in Human Behavior*. Elsevier Ltd, 71, pp. 291–298. doi: 10.1016/j.chb.2017.02.022.

van der Kaa, H. and Krahmer, E. (2014) ‘Journalist versus news consumer: The perceived credibility of machine written news’, *The computation journalism conference New York*, pp. 1–4.

Kari Kelton, Kenneth R. Fleischmann, W. A. W. (2007) ‘Trust in Digital Information’, *Journal of the American Society for Information Science and Technology*, 59(3), pp. 363–374.

Kim, B., Park, J. and Suh, J. (2020) ‘Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information’, *Decision Support Systems*. Elsevier, 134(July 2019), p. 113302. doi: 10.1016/j.dss.2020.113302.

Kind, S. *et al.* (2017) ‘Social Bots | TAB’, (3), p. 84.

Kovach, B. and Rosenstiel, T. (2007) *The Elements of Journalism*. Three Rivers Press: New York.

Kreutzer, R. T. and Sirrenberg, M. (2019) *Künstliche Intelligenz verstehen*, *Künstliche Intelligenz verstehen*. doi: 10.1007/978-3-658-25561-9.

Kunert, M., Hofrichter, J. and SimonAnja, M. (2019) *Glaubwüridigkeit der Medien*. infratest dimap. WDR: Cologne.

Lewis, S. C., Guzman, A. L. and Schmidt, T. R. (2019) ‘Automation, Journalism, and Human–Machine Communication: Rethinking Roles and Relationships of Humans and Machines in News’, *Digital Journalism*. Routledge, 7(4), pp. 409–427. doi: 10.1080/21670811.2019.1577147.

Lokot, T. and Diakopoulos, N. (2016) ‘News Bots’, *Digital Journalism*, 4(6), pp. 682–699. doi: 10.1080/21670811.2015.1081822.

Lucassen, T. and Schraagen, J. M. (2011) ‘Factual accuracy and trust in information: The role of expertise’, *Journal of the American Society for Information Science and Technology*, 62(7), pp. 1232–1242. doi: 10.1002/asi.21545.

Lucassen, T. and Schraagen, J. M. (2012) ‘Propensity to trust and the influence of source and medium cues in credibility evaluation’, *Journal of Information Science*, 38(6), pp. 566–577. doi: 10.1177/0165551512459921.

Mehrabi, D., Hassan, M. A. and Ali, M. S. S. (2009) ‘News media credibility of the internet and television’, *European Journal of Social Sciences*, 11(1), pp. 136–148.

Meier, K. and Reimer, J. (2018) ‘Transparenz im Journalismus. Instrumente, Konfliktpotentiale, Wirkung.’, *„Publizistik. Vierteljahreshefte für Kommunikationsforschung*, (January 2011), pp. 133–155. doi: 10.1007/s11616-011-0116-7.

Metzger, M. J., Flanagin, A. J. and Medders, R. B. (2010) ‘Social and heuristic approaches to credibility evaluation online’, *Journal of Communication*, 60(3), pp. 413–439. doi: 10.1111/j.1460-2466.2010.01488.x.

Mirbabaie, M., Bunker, D., *et al.* (2020) ‘Social media in times of crisis: Learning from Hurricane Harvey for the coronavirus disease 2019 pandemic response’, *Journal of Information Technology*, 35(3), pp. 195–213. doi: 10.1177/0268396220929258.

Mirbabaie, M., Stieglitz, S., *et al.* (2020) ‘Understanding Collaboration with Virtual Assistants – The Role of Social Identity and the Extended Self’, *Business & Information Systems Engineering*, Published. doi: 10.1007/s12599-020-00672-x.

Mittelstadt, B. (2016) 'Auditing for transparency in content personalization systems', *International Journal of Communication*, 10(October), pp. 4991–5002.

Neuberger, C., Langenohl, S. and Nuernbergk, C. (2014) *Social Media und Journalismus*, *Lfm Dokumentation*.

Neuberger, C., Nuernbergk, C. and Rischke, M. (2009) *Journalismus im Internet: Profession - Partizipation - Technisierung*.

Newman, N. (2018) 'Journalism, Media and Technology Trends and Predictions 2018'. Journalism,
Media, And Technology Trends And Predictions 2018. Digital News Report. pp. 2–47.

Plaisance, P. L., Skewes, E. A. and Hanitzsch, T. (2012) 'Ethical Orientations of Journalists Around the Globe: Implications From a Cross-National Survey', *Communication Research*, 39(5), pp. 641–661. doi: 10.1177/0093650212450584.

Preuß, M. *et al.* (2017) 'Fake News und Social Bots – die neuen geheimen Verführer', *Dialogmarketing Perspektiven 2018/2019*, pp. 151–164. doi: 10.1007/978-3-658-25583-1_7.

Sachan, S. *et al.* (2020) 'An explainable AI decision-support-system to automate loan underwriting', *Expert Systems with Applications*. Elsevier Ltd, 144, p. 113100. doi: 10.1016/j.eswa.2019.113100.

Shapiro, A. S. S. and Wilk, M. B. (1965) 'An Analysis of Variance Test for Normality (Complete Samples) Published by : Biometrika Trust Stable URL : http://www.jstor.org/stable/2333709', *Biometrika*, 52(3/4), pp. 591–611.

Shoemaker, P. and Vos, T. (2009) *Gatekeeping theory*. London and New York: Routledge.

Singer, B. J. B. (2006) 'Gate : Newspaper Editors 2004', *J&MC Quarterly*, 83(2).

Singer, J. B. (2014) 'User-generated visibility: Secondary gatekeeping in a shared media space', *New Media and Society*, 16(1), pp. 55–73. doi: 10.1177/1461444813477833.

Stieglitz, S. *et al.* (2019) '"Silence" as a strategy during a corporate crisis – the case of Volkswagen's "Dieselgate"', *Internet Research*, 29(4), pp. 921–939. doi: 10.1108/INTR-05-2018-0197.

Sundar, S. S. (1999) 'Exploring receivers' criteria for perception of print and online news', *Journalism and Mass Communication Quaterly*, 76(2), pp. 373–386.

Taraborelli, D. (2008) 'How the web is changing the way we trust', *Frontiers in Artificial Intelligence and Applications*, 175(1), pp. 194–204.

Tatalovic, M. (2018) 'AI writing bots are about to revolutionise science journalism: we must shape how this is done', *Journal of Science Communication*. Scuola Internazionale Superiore di Studi Avanzati, 17(01). doi: 10.22323/2.17010501.

Thurman, N., Doerr, K. and Kunert, J. (2017) 'Summary for Policymakers', in Intergovernmental Panel on Climate Change (ed.) *Climate Change 2013 - The Physical Science Basis*. Cambridge: Cambridge University Press, pp. 1–30. doi: 10.1017/CBO9781107415324.004.

Thurman, N., Lewis, S. C. and Kunert, J. (2019) 'Algorithms, Automation, and News', *Digital Journalism*. Routledge, 7(8), pp. 980–992. doi: 10.1080/21670811.2019.1685395.

Vázquez-Herrero, J., Direito-Rebollal, S. and López-García, X. (2019) 'Ephemeral Journalism: News Distribution Through Instagram Stories', *Social Media and Society*, 5(4). doi: 10.1177/2056305119888657.

William H. Dutton & Shepherd, A. (2006) 'Trust in the Internet as an experience technology', *Inform Commun Soc*, pp. 433–451.

Wölker, A. and Powell, T. E. (2018) 'Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism', *Journalism: Theory, Practice & Criticism*. doi: 10.1177/1464884918757072.

Paper 8: How Virtuous are Virtual Influencers? – A Qualitative Analysis of Virtual Actors’ Virtues on Instagram

Type (Ranking, Impact Factor)	Conference article (C, N/A)
Status	Published
Rights and permissions	Open Access
Authors	Hofeditz, L., Erle, L., Timm, L., Mirbabaie, M.
Year	2023
Outlet	Hawaii International Conference on System Sciences
Permalink / DOI	https://hdl.handle.net/10125/103051
Full citation	Hofeditz, L., Erle, L., & Timm, L. (2023). How Virtuous are Virtual Influencers? – A Qualitative Analysis of Virtual Actors ’ Virtues on Instagram. Hawaii International Conference on System Sciences. https://hdl.handle.net/10125/103051.

How Virtuous are Virtual Influencers? –A Qualitative Analysis of Virtual Actors' Virtues on Instagram

Conference Paper · January 2023

4 authors:

Lennart Hofeditz
University of Duisburg-Essen
24 PUBLICATIONS 177 CITATIONS
SEE PROFILE

Lukas Erle
Hochschule Ruhr West
2 PUBLICATIONS 0 CITATIONS
SEE PROFILE

Lara Timm
University of Duisburg-Essen
2 PUBLICATIONS 0 CITATIONS
SEE PROFILE

Milad Mirbabaie
Universität Paderborn
132 PUBLICATIONS 1,837 CITATIONS
SEE PROFILE

How Virtuous are Virtual Influencers? – A Qualitative Analysis of Virtual Actors' Virtues on Instagram

Lennart Hofeditz
University of Duisburg-Essen
lennart.hofeditz@uni-due.de

Lukas Erle
University of Duisburg-Essen
lukas.erle@uni-due.de

Lara Timm
University of Duisburg-Essen
lara.timm@uni-due.de

Milad Mirbabaie
Paderborn University
milad.mirbabaie@uni-paderborn.de

Abstract

Recently, virtual influencers (VIs) have become a more frequent alternative to human influencers (HIs). VIs can be described as non-human agents who behave in a human-like pattern. Big enterprises such as Prada, Porsche, Samsung, or Ikea have already collaborated with VIs in the past. Even though it should be clear to users that VIs cannot practice values and virtues in the real world, VIs seem to express certain virtues. This research paper focuses on identifying virtues conveyed by VIs and the effect of expressing virtues on follower engagement by conducting a qualitative content analysis of social media posts. Furthermore, we checked on VIs being abused by companies to convey a more favorable image. Our findings suggest that conveying certain virtues seems to have a positive effect on the engagement. In addition, some VIs were used by companies for virtue signaling without being noticed by their followers.

Keywords: Virtual influencers, virtues, influencer marketing, social media, virtue signaling

1. Introduction

Influencers use social media to promote products, brands, or represent certain opinions in order to win over their followers or steer them in a certain direction (Kádeková & Holienčinová, 2018; Ryan & Jones, 2009). However, influencers do not necessarily have to be real people. For several years now, there has been an increasing emergence of virtual influencers (VIs), who are computer-generated avatars controlled by companies pursuing a specific goal on social media (Arsenyan & Mirowska, 2021). VIs pose lower risks to the public image of companies because they are easier to control and their design is more adaptable (Moustakas et al., 2020). Previous research stated that authenticity and transparency might be important values for human influencers (HIs), but they do not seem to be equally as important for VIs (Robinson, 2020). This is an indication that VIs express different values from those conveyed by HIs. One could think that it would be desirable that VIs mainly show values of good character – commonly referred to as virtues, as suggested by Seligman (2004). However, Vogel et al. (2014) have shown that the use of Instagram can weaken self-esteem if there are upward comparison tendencies. With VIs, this effect could be even stronger as VIs are not bound by limitations of an HI. A too positive expression of virtues displayed by VIs could create upward comparison tendencies. Although previous research already considered different classifications (Batista da Silva Oliveira & Chimenti, 2021), little is known about values and virtues presented by VIs. However, these could be a first indication to better understand the impact of VIs on individuals and society.

Overall, there is a clear deficit in research regarding virtues in the context of social media. This deficit leads to a blurring of moral values (Bowen, 2013), which in turn can harbor another danger: Virtue signaling (VS), which is defined as displaying one's moral values and convictions to the outside world with the aim to convince others of one's moral integrity (Tosi & Warmke, 2016). This can be a problem, as intentionally misleading values and moral intentions can blur actual intentions and weaken trustworthiness and authenticity, which are the most commonly studied characteristics of VIs (Batista da Silva Oliveira & Chimenti, 2021). As VIs are no real human beings, the question arises whether every expressed virtue should be considered VS. However, there is a way to examine the authenticity of virtues in VIs: The parent companies and advertising partners can both be checked for their values and actions. This way, a comparison of the virtues conveyed by VIs with the associated brands and companies can be carried out. To examine virtues expressed by VIs, we pose the following research question (RQ):

RQ1: Which virtues do virtual influencers convey on Instagram and why?

The upward comparison tendencies of followers by too virtuous values expressed by VIs could also lead to a reduction in the number of followers and thus have a negative impact on the company promoting the product. It is therefore important to examine the impact of these virtues on the followers, in order to optimize virtues for the marketing purposes that the VIs were created or hired for (Stapleton et al., 2017). To investigate this impact, we selected engagement as the main dimension of measurement. Engagement can be described as an indication of popularity, trustworthiness and reliability (Batista da Silva Oliveira & Chimenti, 2021). To examine the effect of expressed virtues by VIs on follower engagement, we raise the following second research question:

RQ2: How do virtues conveyed by virtual influencers affect their followers' engagement?

To answer these research questions, we have analyzed 3729 Instagram posts (images and texts) of ten popular VIs and conducted a qualitative content analysis according to Mayring (2015). We deductively considered the virtue framework proposed by Crossan et al. (2013) and compared identified VI virtues with virtues of the VIs' advertisement partners. Our research contributes to e-commerce and marketing research by offering theoretical and practical implications of conveying virtues via VIs.

2. Literature Background

The influencer market has grown enormously in the recent decade and an increasing number of enterprises is cooperating with influencers to expand their reach and popularity. Extensive studies have estimated that brands will invest 15 billion US Dollars in influencer marketing in 2022 (Xie-Carson et al., 2021). Influencers can be defined as people who have social power and affect the habits and thoughts of others. This can happen through spoken and written words, but also through behavior (Robinson, 2020). They have a large number of supporters - named followers - which depend on their opinion regarding fashion, art or lifestyle (Wang et al., 2021).

The research of Freberg et al. (2011) focuses explicitly on social media influencers who have an impact on others through blogs, tweets and social media. VIs share many of these characteristics with their human counterparts, but they also have some additional, unique characteristics: They are embodied virtual agents with digital avatars, which makes them seem tangible and realistic (Arsenyan & Mirowska, 2021; Tan & Liew, 2020). Furthermore, VIs can be designed using both CGI and AI (Moustakas et al., 2020). In contrast to social bots (Stieglitz et al., 2022) or automated news accounts (Hofeditz et al., 2021), they are not automated but manually controlled by humans and represent a certain character with corresponding behavioral patterns (Arsenyan & Mirowska, 2021; Najari et al., 2021). A more comprehensive overview of the different virtual entities can be found in Table 1 below:

Table 1. Description of virtual entities

	Definition
VI	"agents augmented with digital avatars, designed to look human" (Arsenyan & Mirowska, 2021, p.2), mostly controlled by humans
CGI Influencer	Subcategory of VIs, which are computer-generated individuals who have real human traits, characteristics and personalities (Moustakas et al., 2020; Sobande, 2021; Xie-Carson et al., 2021)
AI Influencer	Subcategory of VIs, which are based on algorithms and machine learning in order to perform like a real person (Kumar et al., 2019; Moustakas et al., 2020)
Virtual Avatar	Images of persons, which are entirely controlled by the users (von der Pütten et al., 2010)
Social Bot	"a computer-based algorithm that automatically controls a social media account, produces content, and potentially interacts with human users on social media trying to emulate human behavior" (Najari et al., 2021, p.1)

As existing definitions of VIs slightly differ, we have combined multiple sources into the following, summarizing definition: VIs are non-real characters (human-like or non-human) designed for either marketing purposes or to simply create engagement on social media. They have inherent social power to shape the behavior of others through their own words and actions. Moreover, VIs can be categorized not only by the nature of their construction, but also by their appearance; thus, one can identify VIs, which are either humanlike or cartoonlike (Xie-Carson et al., 2021).

For companies and brands, VIs are especially important, because people have a greater trust in unique, human-like influencers conveying values and virtues (Batista da Silva Oliveira & Chimenti, 2021). One of the most significant advantages for companies is that VIs are not real people with a free will, meaning that publicity risks can be avoided (Xie-Carson et al., 2021). For example, VIs are less likely to be involved in scandals as enterprises have a much greater control over their content and presentation (Arsenyan & Mirowska, 2021; Moustakas et al., 2020). Beyond that, "unlike HIs, where their personal life choices may affect the perception of the brand they promote, VIs are ageless human robots who do not have an 'offline life' which could negatively affect their 'online persona'" (Moustakas et al., 2020, p.2**)**. For cooperating brands it is important that influencers are credible, attractive, and trustworthy to create a higher level of engagement consisting of metrics like comments, interactions, or likes (Batista da Silva Oliveira & Chimenti, 2021; Djafarova & Rushworth, 2017). Engagement can improve follower bonding and generates a higher coverage for advertised products. VIs further generate their influence through the attractiveness stereotype, their human likeness, and audio-visual effects (Faddoul & Chatterjee, 2020; Khan & Sutcliffe, 2014). To create unique characters, VIs are equipped with their own personality traits which are often expressed by values and virtues, such as solidarity with the black lives matter movement (Hofeditz et al., 2022).

3. Theoretical Background

Virtues are ''acquired human qualities, the excellences of character, which enable a person to achieve the good life'' (Mintz 1996, p. 827) and are understood as intrinsic qualities. Peterson and Seligman (2004) deduce from the main doctrines of virtues according to Aristotle and Plato that the fundamental virtues applied in historical literature also fit the description of a good character. These virtues are Wisdom, Courage, Temperance, Justice, Transcendence, and Humanity. Aristotle further describes a virtue as a moral quality that is influenced by an individual's actions (Aristotle, n.d.). The core concept of his elaboration aims at the selection of a decision that lies between the two extremes of a virtue – deficiency and excess. Virtues are characteristics that usually evoke delight or sorrow in the observer (Tasset, 2019). For example, for *Humanity*, possible manifestations of the two extremes are: *harsh/cruel* on the excess side and *obsequious* on the deficiency side, with the virtous mean being *kindness*. Thusly, virtues describe virtuous qualities that distinguish a character, and they are shaped by personal decisions and made visible to the rest of the world. These qualities of virtues were summarized in a framework for ethical decision making by Crossan et al. (2013).

On social media, virtues can be amplified by portraying an idealized self (Wallace et al., 2020). However, not only good values are conveyed on social media: A popular example for non-virtuous behavior is popular TikTok influencer Andrew Tate, who has recently been banned by the platform for promoting sexist values to a large audience (Cooper, 2022). In addition to this, the phenomenon of VS can occur: According to Tosi and Warmke (2016), VS is an active contribution to moral discourse with the intention of convincing others of one's own moral integrity. VS can also be used with the intent to mislead, when a virtue or virtuous image is intentionally created in a dishonest manner to signal an image that is different from reality (Levy, 2021). An example for dishonest VS would be falsely pretending to be concerned about environmental disasters in order to increase one's own moral defensibility and social acceptance.

As VIs are no real humans and virtues are based on intrinsic attitudes, they cannot really be virtuous. However, they express certain values and virtues on their social media channels which may be an indication for VS to promote products of partnering brands. As this would be an ethical issue and little is known about the values and virtues of VIs, we further examined VIs' social media posts and compared it with values of partnering brands.

4. Research Design

In order to examine virtues conveyed by VIs on social media, we conducted a qualitative content analysis (Mayring, 2015) on popular VIs. Furthermore, we considered the engagement for posts where they expressed certain values. Lastly, we compared these virtues with values of companies partnering with the VIs.

4.1 Choice of Influencers

Firstly, we selected the most engaging VIs as those had a high reach and multiple partnerships with brands and companies. This process of selecting suitable VIs is summarized in Figure 1.

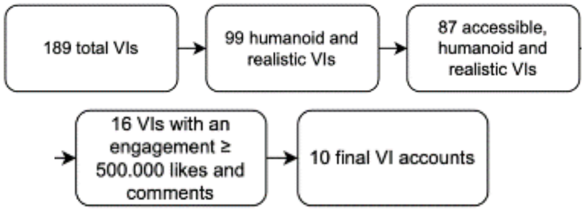

189 total VIs
99 humanoid and realistic VIs
87 accessible, humanoid and realistic VIs
16 VIs with an engagement ≥ 500.000 likes and comments
10 final VI accounts

Figure 1. VI selection process

Because the present research aims to draw connections between HIs and VIs, choosing comparable VIs is sensible. These humanoid VIs should furthermore be realistic in their appearance. The most comprehensive and complete list of VIs is offered by the website VirtualHumans.org (2021). Out of 189 total VIs listed on their database, a total of 99 accounts featured realistic and humanoid characters. Not all accounts were public which led to twelve VIs having to be excluded.

All VIs were then compared to each other with regards to their posts' overall engagement. Follower counts were disregarded as they might be inflated by inactive or fake accounts (Instasize, 2021; Logan, 2018). Furthermore, our second research question directly addresses engagement, further strengthening the importance of this metric. The accounts with the largest engagement were further checked for whether they used image captions at all and whether these captions were in English. This parameter eliminated four VI accounts who either did not use captions or used captions in another language. Another account was excluded from the analysis as it was transformed into a company blog for a fashion retailer, featuring many normal humans. A sixth account was also excluded for repeatedly posting surreal images and seldomly featuring the VI itself.

Adapted from Alibakhshi and Srivastava (2022), we measured the overall engagement by combining the total number of likes and comments. Influencers with a total engagement of more than 500.000 likes and comments can be defined as mega- and macro-influencers with the biggest reach and influence (mediakix, 2021). Therefore, all VI accounts with a total engagement above this threshold were considered for the final analyses. This led to a total of ten VIs. These are (in order of their total engagement, highest to lowest): Miquela Sousa (@lilmiquela), Imma (@imma.gram), Rozy (@rozy.gram), Leya Love (@leyalovenature), Zinn (@plusticboy), Ion Göttlich (@iongottlich), Blawko (@blawko22), Shudu (@shudu.gram), Binxie (@itsbinxie), and Bermuda (@bermudaisbae).

4.2 Codebook and Content Analysis

As this paper aims to examine which virtues are portrayed and communicated by VIs, we conducted a content analysis according to Mayring (2015). We decided to use a blend of both inductive and deductive content analysis steps: First we defined categories and keywords by building on Crossan et al. (2013) and considering similar literature. The VIs' posts were also examined regarding additional keywords and categories.

To properly and reliably code the content extracted from the VIs' Instagram profiles, we developed a codebook containing keywords and categories. As a foundation for the categories we used the framework of Crossan et al. (2013) as a basis. The different expressions of each category provided an initial collection of possible keywords.: Each virtue contains a set of three to four words describing the Deficiency, Mean and Excess. An example would be the virtue Courage, offering the keywords *cowardice, laziness* and *inauthenticity* for its Deficiency expression, the words *bravery, persistence,* and *integrity* for its Mean expression and finally the words *recklessness, zealot* and *righteousness* for its Excess expression. Some words were seldomly used in the modern language used by influencers on social media (Eisenstein et al., 2014). To counteract this circumstance, we used the dictionary database Thesaurus, and added suitable synonyms to the codebook, which were used more frequently. In some cases, we further decided to add sensible additional words that describe the literature-derived keywords in more detail. An example for this would be the keyword *creativity* from the virtue Wisdom: When influencers talk about their creative activities, they rarely use the word "creativity" itself, but rather talk about the exact activity being part of creativity. Hence, we added activities like singing or dancing as keywords, which in turn represent the keyword creativity. To allow the lexical search to identify all posts belonging to a category, some words have been stemmed: For example, the keyword *Creativity* itself would only identify posts containing this exact wording. Instead, we used the keyword *creat**, which is able to identify more words like *Creativity*, *Create*, and more.

A further important step in creating the codebook included the creation of a seventh coding category, which was named Brand Partnership and aimed at coding posts that were sponsored or paid for by any brand. The keywords for this category were derived from a report from the British Advertising Standards Authority (2019). These keywords were again stemmed and extended by the words story and bio, as per experience sponsorships can often be identified by influencers linking brands or products in either their Instagram page bio or their timed stories. Table 2 is an excerpt of our codebook, offering a collection of example words per category.

Table 2. Codebook excerpt

Wisdom	*Courage*	*Humanity*
creat* (creativity)	protest*	queer*
curious* (curiosity)	riot*	tolera* (tolerance)
learn* (learning)	authent* (authenticity)	free* (freedom)

To further analyze the posts, we added associated emoticons. For instance, the word *free** is linked to emoticons like 🕊️, ☮️, or ✌️. This codebook was then tested via an intercoder reliability analysis to ensure all coders would apply codes the same way. The sample was extracted by setting a random time frame of three months and included 250 individual posts. These posts were coded according to the codebook. In the end, 969 codes were assigned by four coders, with some posts being assigned to multiple categories. This resulted in a Fleiss Kappa of κ = .994 (Fleiss & Cohen, 1973).

4.3 Analysis Tools and Steps

We selected Instagram as a suitable social media platform because it is one of the platforms with the most active user bases for both VIs and content marketing and telling stories and creating personalities works best with pictures (Faßmann & Moss, 2016). Advertising on Instagram increases followers' willingness to purchase, enabling the platform to combine marketing with IS (Qiu et al., 2021). Finally, image-based social media networks generally serve more diverse purposes to their users than simple text-based networks like Twitter (Teo et al., 2019). However, Instagram does not allow a direct export of post and account data. Therefore we used the social media analytics tool Fanpage Karma[1]. The ten VIs were added to a dashboard to allow an overview of all posts over a set time period which has been set from November 1st, 2011 (the earliest accessible date) to the day the analysis was started (November 30th, 2021). This resulted in a total of 3729 posts across all ten VIs. For further analysis and coding of these posts, the tool MaxQDA[2] was used. Firstly, exported data was uploaded into the tool. Secondly, a lexical search for the keywords was conducted to check all posts for these keywords and remove posts that featured only parts of the keywords or used them in a different context. Finally, these keywords were auto coded by matching the keywords with the respective category code label we had defined beforehand. To explore possible new virtues, a word cloud of the most frequent words was created. If these words matched one of the already existing categories but had not been included in the codebook yet, it was added to the respective category. Alternatively, if a word did not match any of the existing categories, it was noted down and later clustered and grouped into new categories by considering the literature mentioned in the previous sections. Finally, the new categories were added as codes to MaxQDA and the posts were auto-coded again.

While MaxQDA only supports the semi-automated coding of caption as part of its text processing capabilities, the coding of images was done manually. The image labeling was done by checking the posts for the visual representation of the keywords. For example, if the post included a religious sign like a Star of David, it was sorted into the category *Transcendence* as it represents the keywords *faith* and *religion*. As another example, if a picture showed the VI meeting up with friends, the picture was sorted into the category *Humanity*, as it showed a visual representation of the keywords *friends* and *togetherness*. These visual representations were noted down in a spreadsheet containing the individual post, the keywords that were identified, and the visual representation.

Finally, the combination of virtues conveyed – both visually and in the captions – provided information about which virtues the VIs convey. Adressing the second research question, after having conducted all steps mentioned above, the average and peak engagement metrics of all VIs were measured using MaxQDA. The average engagement of a VI's posts matching our categories was then compared against that VI's average engagement and the average engagement of posts that did not match any category. This analysis was conducted for each VI individually and for all VIs

[1] https://www.fanpagekarma.com/de

[2] https://www.maxqda.de

together separately. The VI with the fewest total posts was Shudu with just 99 posts in total. In order to be able to compare a similar number of posts for each VI, we settled on comparing the most recent 100 posts of each VI, not least to be able to calculate percentages more easily.

5. Results

Each main virtue category of Crossan et al. (2013) was found in the analysis of the ten VIs' posts. Different frequencies resulted for the individual virtues[3]. Virtues related to the perception of nature, *Humanity* and spirituality were most frequently used, whereas virtues such as *Courage* or *Justice* were rarely represented. The virtues mainly conveyed by VIs are *Humanity, Wisdom, Transcendence, Temperance* and *Courage*. These virtues were often displayed through posts about interpersonal relationships like meeting friends or interacting with pets, posts about nature and environment, artistic outlets and spiritual topics, as well as openly talking about own emotions – both positive and negative ones. Most importantly, the assessment of the VIs' posts did not result in any additional, new virtues.

To check how the conveyed virtues – if at all – have an influence on the followers' engagement, the last 100 posts of each influencer were analyzed. First, the average engagement of each influencer was calculated from the existing posts. The posts were then divided into three categories: Posts that conveyed at least one virtue, posts that exceeded the average individual engagement of the VI's profile, and the combination of both. More than 80% of the posts with a higher-than-average engagement conveyed at least one virtue. From this we deduce that posts in which virtues are represented generate a higher level of engagement than posts in which no virtues are conveyed.

Furthermore, we considered all posts coded as Brand Partnership that also conveyed a virtue to analyze whether these virtues aligned with what the advertisement partners stand for, or whether VS might have occurred. To this end, we first listed all companies the VI had partnered with over the course of their last 100 posts and noted which individual virtues the sponsored posts conveyed. We then systematically searched for the company name in combination with keywords such as "controversy" and "scandal", as well as researching information regarding the company itself. This led to news articles either condemning or praising certain actions or comments by the company in the past, which we used to decide whether the results contradicted with any of the virtues conveyed by the VIs' posts. Information given by companies themselves – for example as part of statements or product descriptions – also helped with this assessment, as claims of sustainability or good worker conditions could be easily researched by consulting independent sources. When conducting this examination of companies' virtues, it was imperative to only focus on post-specific virtues: For example, if a company was involved in a sexism scandal, yet none of the sponsored posts that company had with a VI conveyed *Humanity*, this scandal would not be considered as VS.

A total of 64 companies had partnered with the chosen VIs. One positive example in which a post's virtues and the corresponding company's virtues align can be found in the partnership between the VI Rozy and the brand Maison Margiela Fragrances: Rozy's post conveys the virtues *Temperance*, *Humanity* and *Courage*. These virtues align well with the fact that the brand offers many different unisex fragrances and its fashion label – Maison Margiela – has been praised for offering genderless clothing collections (Lim, 2021). In contrast, an example for dishonest VS was found in sponsored posts of the VI Shudu and the jewelry brand Tiffany & Co. The posts by Shudu, in which she can be seen modelling with diamonds provided by Tiffany & Co., convey the virtues *Transcendence*, *Wisdom* and – most importantly – *Humanity*. In contrast to this last virtue, Tiffany & Co. has repeatedly been accused of using so called blood diamonds for their jewelry. These diamonds are often unethically sourced or "originate from mines that employ slave labor systems" (Osmond, 2021, para. 4). These sponsored posts can therefore be seen as malicious VS. A notable example outside of the norm is the partnership between the VI Shudu and the Italian fashion label Ferragamo. In this case, Shudu's posts modelling for this brand conveyed the virtues *Transcendence* and *Wisdom*. Important for our analysis however was the fact that Shudu is a model of color and shortly before the partnership between her and Ferragamo, the brand was involved in a racism scandal as employees alleged racist treatment of both customers and employees within the stores, as well as racist remarks form high-ranking representants of the brand during a photoshoot with models of color (Alleyne,

[3] The total number of virtues found during the analysis can be found here: https://tinyurl.com/2p9ddt9f3

2020; Barry, 2020). While racist remarks do not necessarily contrast with the two virtues conveyed by the posts, a model of color partnering with a fashion label facing racism allegations might be considered dishonest VS.

Generally, the majority of sponsored posts can be considered virtue-congruent with only 13 out of 64 partnerships raising points of concern. However, all posts conveying *Humanity* that feature a partnership with a fashion label that produces its clothes in the middle east or Asia could be considered as VS, as the workers' rights and working conditions cannot be properly assessed.

6. Discussion

The by far most frequently expressed virtue among the examined VIs was *Humanity*. The virtue was often expressed through posts about interpersonal relationships. This suggests that operators of VIs consider *Humanity* to be highly important in shaping a VI's character. However, previous research has found evidence that a large proportion of viewers of VIs do not perceive them as humanlike (Hofeditz et al., 2022). Especially with regard to negative effects of the associated upward comparison tendencies on social media (Vogel et al., 2014), this raises new ethical challenges and questions regarding the comparison with VIs and the effects on self-esteem.

By conveying virtues such as *Humanity*, *Wisdom*, *Transcendence* or *Courage* (which were the most frequently expressed ones) VIs seem to live similar lives to their followers: They meet friends, own pets, show strong emotions, and have artistic outlets. Most of the time, these postings do not contain any specific message and exist as stand-alone content. This can be put into the context of the work of Mintz (1996), who claims that virtues are a central part of our human self and are the reason we function the way we do. Virtues could be used as metaphorical masks by VIs to pretend to be living a valuable human life, which in turn could be one of the reasons users show such interest in this content.

Some VIs such as Leya Love focus on few virtues: Even though Leya Love portrays a variety of virtues, *Transcendence* is her focus: She covers topics like mental health, meditation, and environmental issues and acts as an educator and advocate on topics like the ongoing climate crisis and endangered animals. Ion Göttlich serves a similar purpose: He offers information on biking sport, which is his exclusive content since he is strongly involved with a bike suppliance company. This supports the arguments made by Robinson (2020): As long as the content appeals to the audience and they can relate to what the VIs seemingly experience in their everyday life, there is no need for a physical existence. The story being told is far more important than the realness, which contradicts with social media users' need for more authenticity (Robinson, 2020).

Moreover, the results show that posts that represent virtues generate a higher engagement. It seems that humans not only prefer human-realistic avatars over less-realistic ones – as shown by Seymour et al. (2021) – but also like to see virtues expressed by VIs such as *Humanity* and *Wisdom*. Even though taking advantage of virtue portrayal seems to increase the total number of interactions with the posts of the VIs, it should be well planned: Companies need to investigate if the virtues conveyed by the VIs' posts are congruent with their own virtues, since this could lead to controversy. As perfect as VIs seem to be due to the lack of impulsive reactions (Robinson, 2020), it is even more important for companies hiring a VI to look deeper into their own past scandals and problematic behavior.

Ethical implications, as proposed by Crossan (2013) are difficult to gauge: The VIs seem to behave like classic, HIs for the most part and show similar attributes and virtues both in previous research (Moustakas et al., 2020), and in the present study. Since VIs cannot make human mistakes – like reacting impulsively or emotionally – VIs are less prone to be involved in scandals (Robinson, 2020). This implies that it is very important for both clients and the influencer's companies to make sure the virtues of the post are not going to cause any controversy.

Overall, it seems like VIs largely do not differ from HIs in their virtuosity. VIs also serve human attributes like "talent, beauty, style, comedy, sensuality, or authority" (Batista da Silva Oliveira & Chimenti, 2021). Their content covers very similar topics like Art, Beauty, Fashion, LGBTQ awareness, Lifestyle and more (Rundin & Colliander, 2021). Faddoul and Chatterjee (2020) state that as long as the presented framework satisfies the requirements for humans to accept the virtual entity as a valid personality, VIs can serve as a substitute to a real influencer. This is supported by our findings. However, the authors also stress that emotional factors should not be excluded, which would mean that a bigger focus should be set on virtues like *Temperance*, a virtue that is not as present in the posts as the virtue *Humanity*. Lastly, it is important to note that even though the attractiveness of the VI pays a big role when it comes to persuasiveness

(Khan & Sutcliffe, 2014), it has to be treated with caution: Research has shown that the audience desires more authenticity in terms of less perfect influencers (Osburg & Heinecke, 2019). An appearance that is too flawless and perfect can create distrust and pushes consumers away, rather than attracting them. When VIs become more realistic, adding natural flaws to their appearance (like the gap between Miquela's teeth) or personality is a way to circumvent this distrust.

Our analysis identified 13 posts that we considered concerning in terms of VS, as suggested by Levy (2021). VIs can be used more easily to disguise unethical behavior, as they are unable to provide feedback to the enterprises and both their content and appearance can be manipulated easily. HIs usually hold their own opinions and can express their point of view. Before agreeing to an advertisement deal, they can inform themselves regarding the client company and consider whether their approach is in line with their own values. By using VIs, companies can convey certain virtues and ensure that their products are associated with values that HIs would not be willing to convey. We further found no evidence that the accounts of dishonest VS carried out by VIs was recognized as such among their followers. Thus, conveying virtues through VIs seems to be a good opportunity for a company to paint a more positive image of themselves. While the companies behind the VIs could also reject questionable partnerships, this could be easily circumvented if brands create their own VIs. For consumers, it is therefore important to critically question the virtues conveyed in advertisements posted by VIs.

However, VS can also be used positively (Levy, 2021), for instance to emphasize the commitment to sustainability or human rights. This way, a brand can communicate its ethical principles to its customers and generate engagement through fair working conditions or environmental protection. Furthermore, previous research revealed a relationship between VS and offline behavior intention (Wallace et al., 2020). It is possible, then, that conveying virtues through VIs may not only lead followers to internalize them - which is beneficial for society - but could also have a positive impact on companies' offline actions when they collaborate with VIs that convey virtues.

7. Limitations and Future Research

There are some limitations to this paper. First, the selection of influencers does not represent all VIs: This paper focused on the VIs that generate the highest engagement on their profiles. This could lead to distorted results, as VIs with less engagement could have different characteristics or maybe convey virtues either very differently or not at all. Further research should also include VIs with less followers and engagement.

In addition, it is important to state that the analyzed VIs form a heterogeneous group, yet environmental activist VIs may generate different engagement than fashion VIs. These differences should be considered in future research.

We recommend that future research should focus more on how and why exactly people get involved with VIs and what this means in terms of ethical challenges for non-human personalities. When do VIs become too human and what threat can they pose? What opportunities do they hold? Can artificial intelligence be implemented or is it safer to keep them controlled by a company? Observing and understanding how and why the followers react to VIs and how open they are accepting their existence in their everyday life could offer answers to these questions. As we exclusively coded the images and tags of VI's Instagram posts to derive our results, future research should examine the comments on the posts. Furthermore, we recommend running a similar analysis of values and virtues conveyed by HIs and compare these to our findings to establish a better comparability between VIs and HIs.

8. Conclusion

This study has made several contributions to research on the topic of virtuosity and ethics of VIs: Our findings suggest that VIs convey *Humanity* as the most frequently expressed virtue (followed by *Wisdom*, *Transcendence* and *Temperance*). In addition, conveying virtues seems to positively influence followers' engagement. Furthermore, the most frequently expressed virtues are the ones to most likely contain VS. There have been a handful of partnerships which signaled virtues that do not align with the values and ethics of the companies being advertised. This is problematic, as VS can decrease the followers' trust in both the VI and the company. We found that there is evidence of VIs that have been used for virtue signaling by companies, without their followers taking note.

In total, we contribute to IS research by providing knowledge on how VIs express values and virtues and how they are used to increase their engagement. This implies that first predictions on the impact of ethical behavior on follower engagement (and subsequently marketing value) can be established. We examined how VS can be used by companies to convey more positive values in combination with a product without being noticed. However, the communication of virtues by VIs can not only result in dishonest VS and a reduced self-esteem of the followers due to an upward comparison, but also possibly lead to followers and companies acting more virtuously.

We contribute to marketing practice by offering practitioners more insight into the way in which VIs can be used to convey certain virtues and values in their posts. This knowledge can be used to choose VIs that align with a company's values – or the ones they want to convey – for sponsorships and advertisement campaigns. However, our findings also warn consumers that virtue signaling can occur with VIs.

9. References

Alibakhshi, R., & Srivastava, S. C. (2022). Post-Story: Influence of Introducing Story Feature on Social Media Posts. Journal of Management Information Systems, 39(2), 573–601.

Alleyne, A. (2020). The fashion industry says it stands against racism. Critics aren't buying it. *CNN Style*. https://edition.cnn.com/style/article/fashion-industry-black-lives-matter/index.html

Aristotle. (n.d.). *Nicomachean Ethics*. Harvard University Press. Abgerufen 6. Dezember 2021, von http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.01.0054

Arsenyan, J., & Mirowska, A. (2021). Almost human? A comparative case study on the social media presence of virtual influencers. *International Journal of Human-Computer Studies*, *155*, 102694.

Barry, C. (2020). Luxury fashion brands forced to confront racism in the industry. *PBS News Hour*. https://www.pbs.org/newshour/arts/luxury-fashion-brands-forced-to-confront-racism-in-the-industry

Batista da Silva Oliveira, A., & Chimenti, P. (2021). „Humanized Robots": A Proposition of Categories to Understand Virtual Influencers. *Australasian Journal of Information Systems*, *25*.

Bowen, S. A. (2013). Using Classic Social Media Cases to Distill Ethical Guidelines for Digital Engagement. *Journal of Mass Media Ethics*, *28*(2), 119–133.

Cooper, C. (2022). The values of 'Manhood': The dangerous rise of Andrew Tate. The Oxford Blue. https://www.theoxfordblue.co.uk/2022/08/25/the-values-of-manhood-the-dangerous-rise-of-andrew-tate/.

Crossan, M., Mazutis, D., & Seijts, G. (2013). In Search of Virtue: The Role of Virtues, Values and Character Strengths in Ethical Decision Making. *Journal of Business Ethics*, *113*(4), 567–581.

Djafarova, E., & Rushworth, C. (2017). Exploring the credibility of online celebrities' Instagram profiles in influencing the purchase decisions of young female users. *Computers in Human Behavior*, *68*, 1–7.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of Lexical Change in Social Media. *PLoS ONE*, *9*(11), e113114.

Faddoul, G., & Chatterjee, S. (2020). A Quantitative Measurement Model for Persuasive Technologies Using Storytelling via a Virtual Narrator. *International Journal of Human–Computer Interaction*, *36*(17), 1585–1604.

Faßmann, M., & Moss, C. (2016). *Instagram als Marketing-Kanal*. Springer Fachmedien Wiesbaden.

Fleiss, J. L., & Cohen, J. (1973). The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement*, *33*(3), 613–619.

Freberg, K., Graham, K., McGaughey, K., & Freberg, L. A. (2011). Who are the social media influencers? A study of public perceptions of personality. *Public Relations Review*, *37*(1), 90–92.

Hofeditz, L., Mirbabaie, Mi., Stieglitz, S., & Holstein, J. (2021). Do You Trust An AI-Journalist? A Credibility Analysis Of News Content With AI-Authorship. *Proceedings of the European Conference of Information Systems*. Wellington, NZ.

Hofeditz, L., Nissen, A., Schütte, R., & Mirbabaie, M. (2022). Trust Me, I'm An Influencer! - A Comparison Of Perceived Trust In Human And Virtual Influencers. *Proceedings of the European Conference of Information Systems*. Timisoara, Romania.

Instasize. (2021). Increase Engagement by Removing Instagram Ghost Followers. *instasize*. https://instasize.com/blog/increase-engagement-by-removing-instagram-ghost-followers

Kádeková, Z., & Holienčinová, M. (2018). Influencer Marketing as a Modern Phenomenon Creating a New Frontier of Virtual Opportunities. *Communication Today*, *9*(2). http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.cejsh-29ccb3e1-8052-4f82-b9f7-b61b1a58fe21

Khan, R. F., & Sutcliffe, A. (2014). Attractive Agents Are More Persuasive. *International Journal of Human-Computer Interaction*, *30*(2), 142–150.

Kumar, V., Rajan, B., Venkatesan, R., & Lecinski, J. (2019). Understanding the Role of Artificial Intelligence in Personalized Engagement Marketing. *California Management Review*, *61*(4), 135–155.

Levy, N. (2021). Virtue signalling is virtuous. *Synthese*, *198*(10), 9545–9562.

Lim, J. (2021). Maison Margiela unveils genderless Icons Collection for AW21. *The Industry Fashion*. https://www.theindustry.fashion/maison-margiela-unveils-genderless-icons-collection-for-aw21/

Logan, M. (2018). Is the Follower Count Dead? Why Bloggers Need to Look Beyond Reach & Focus on This Key

Metric…. *medium*. https://medium.com/willu/is-the-follower-count-dead-why-bloggers-need-to-look-beyond-reach-focus-on-this-key-metric-21fc04395de6

Mayring, P. (2015). Qualitative Content Analysis: Theoretical Background and Procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Hrsg.), *Approaches to Qualitative Research in Mathematics Education* (S. 365–380). Springer Netherlands.

mediakix. (2021). Influencer Tiers For The Influencer Marketing Industry. *Mediakix*. https://mediakix.com/influencer-marketing-resources/influencer-tiers/

Mintz, S. M. (1996). Aristotelian virtue and business ethics education. *Journal of Business Ethics*, *15*(8), 827–838.

Moustakas, E., Lamba, N., Mahmoud, D., & Ranganathan, C. (2020). Blurring lines between fiction and reality: Perspectives of experts on marketing effectiveness of virtual influencers. *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, 1–6.

Najari, S., Salehi, M., & Farahbakhsh, R. (2021). GANBOT: A GAN-based framework for social bot detection. *Social Network Analysis and Mining*, *12*(1), 4.

Osburg, T. H., & Heinecke, S. (Hrsg.). (2019). *Media trust in a digital world: Communication at crossroads*. Springer.

Osmond, P. (2021, September 14). Tiffany & Co. Creates controversy with new campaign. *The Tufts Daily*. https://tuftsdaily.com/arts/2021/09/14/tiffany-co-creates-controversy-with-new-campaign/

Peterson, C., & Seligman, M. (2004). *Character Strengths and Virtues: A Handbook and Classification*. Oxford University Press.

Qiu, L., Chhikara, A., Vakharia, A. (2021) Multidimensional Observational Learning in Social Networks: Theory and Experimental Evidence. Information Systems Research 32(3):876-894.

Robinson, B. (2020). Towards an Ontology and Ethics of Virtual Influencers. *Australasian Journal of Information Systems*, *24*.

Rundin, K., & Colliander, J. (2021). Multifaceted Influencers: Toward a New Typology for Influencer Roles in Advertising. *Journal of Advertising*, *50*(5), 548–564.

Ryan, D., & Jones, C. (2009). *Understanding digital marketing: Marketing strategies for engaging the digital generation*. Kogan Page.

Seymour, M., Yuan, L. (Ivy), Dennis, A. R., & Riemer, K. (2021). Have We Crossed the Uncanny Valley? Understanding Affinity, Trustworthiness, and Preference for Realistic Digital Humans in Immersive Environments. *Journal of the Association for Information Systems*, *22*(3), 591–617.

Sobande, F. (2021). Spectacularized and Branded Digital (Re)presentations of Black People and Blackness. *Television & New Media*, *22*(2), 131–146.

Stapleton, P., Luiz, G., & Chatwin, H. (2017). Generation Validation: The Role of Social Comparison in Use of Instagram Among Emerging Adults. *Cyberpsychology, Behavior and Social Networking*, *20*(3), 142–149.

Stieglitz, S., Hofeditz, L., Brünker, F., Ehnis, C., Mirbabaie, M., & Ross, B. (2022). Design principles for conversational agents to support Emergency Management Agencies. International Journal of Information Management, 63, 102469.

Tan, S.-M., & Liew, T. W. (2020). Designing Embodied Virtual Agents as Product Specialists in a Multi-Product Category E-Commerce: The Roles of Source Credibility and Social Presence. *International Journal of Human–Computer Interaction*, *36*(12), 1136–1149.

Tasset, J. L. (2019). Bentham on ‘ Hume ’ s Virtues ’. In G. Varouxakis & M. Philip (Eds.), Happiness and Utility (pp. 81–97). UCL Press. http://www.jstor.com/stable/j.ctvf3w1s5.9.

Teo, L. X., Leng, H. K., & Phua, Y. X. P. (2019). Marketing on Instagram: Social influence and image quality on perception of quality and purchase intention. *International Journal of Sports Marketing and Sponsorship*, *20*(2), 321–332.

The Advertising Standards Authority. (2019). *The labelling of influencer advertising A report on what labels and other factors help people understand when influencers are posting advertising content*. https://www.asa.org.uk/static/uploaded/e3158f76-ccf2-4e6e-8f51a710b3237c43.pdf

Tosi, J., & Warmke, B. (2016). Moral Grandstanding. *Philosophy & Public Affairs*, *44*(3), 197–217.

VirtualHumans. (2021). About VirtualHumans.org | Create a Virtual Influencer. *VirtualHumans.Org*. https://www.virtualhumans.org/about

Vogel, E. A., Rose, J. P., Roberts, L. R., & Eckles, K. (2014). Social comparison, social media, and self-esteem. *Psychology of Popular Media Culture*, *3*(4), 206–222.

von der Pütten, A. M., Krämer, N. C., Gratch, J., & Kang, S.-H. (2010). “It doesn’t matter what you are!” Explaining social effects of agents and avatars. *Computers in Human Behavior*, *26*(6), 1641–1650.

Wallace, E., Buil, I., & de Chernatony, L. (2020). ‘Consuming Good’ on Social Media: What Can Conspicuous Virtue Signalling on Facebook Tell Us About Prosocial and Unethical Intentions? *Journal of Business Ethics*, *162*(3), 577–592.

Wang, Huang, Q., & Davison, R. M. (2021). How do digital influencers affect social commerce intention? The roles of social power and satisfaction. *Information Technology & People*, *34*(3), 1065–1086.

Xie-Carson, L., Benckendorff, P., & Hughes, K. (2021). Fake it to make it: Exploring Instagram users’ engagement with virtual influencers in tourism. *Travel and Tourism Research Association: Advancing Tourism Research Globally*, *17*.