universität ulm

Responsible Digital Technology
–
Contributions to Foster Explainable Artificial Intelligence and Empower Marginalized Groups

Dissertation
zur Erlangung des Grades
einer Doktorin der Wirtschaftswissenschaften (Dr. rer. pol.)
eingereicht an der
Fakultät für Mathematik und Wirtschaftswissenschaften
der Universität Ulm

vorgelegt von
Irina Sigler
M. Sc. Wirtschaftswissenschaften
geb. in Tjumen (Russland)

Gutachter
Prof. Dr. Mathias Klier
Prof. Dr. Steffen Zimmermann

2023

Amtierender Dekan der Fakultät:

Prof. Dr. Stefan Funken

Gutachter:

Prof. Dr. Mathias Klier
Prof. Dr. Steffen Zimmermann

Promotionskolloquium:

22.03.2023

Acknowledgments

I want to express my deepest gratitude to Prof. Dr. Mathias Klier and Prof. Dr. Julia Klier for their exceptional expertise, guidance, and support throughout this research project. Your invaluable feedback pushed my research to a higher level.

I also want to thank the Institute of Business Analytics team for our successful collaboration and the inspiring team atmosphere. Special thanks go to my co-authors Maximilian Förster, Kilian Kluge, Katharina Kaufmann, Julia Brasse, and Hanna Broder.

Finally, I could not have completed this dissertation without the support of my parents, Valentina and Vladimir Hardt, and my husband, Christian Sigler, who I love.

Thank you.

Irina Sigler

Summary of Contents

Table of Contents	II
List of Figures	III
List of Tables	IV
List of Abbreviations	V
1. Introduction	1
2. Research on Digital Technologies to Alleviate Societal Challenges	40
3. Research on Societal Challenges Posed by Digital Technologies	95

Explanatory note: *Each paper in this dissertation is presented as an individual manuscript concerning figures, tables, general numbering, and references and is thus self-contained to facilitate selective reading.*

Table of Contents

List of Figures III

List of Tables IV

List of Abbreviations V

1. Introduction 1

    1.1 Motivation 1

    1.2 Research Objectives 7

    1.3 Research Paradigms and Research Methods 16

    1.4 Structure of the Dissertation 22

    1.5 References 23

2. Research on Digital Technologies to Alleviate Societal Challenges 40

    2.1 #JOBLESS #OLDER #DIGITAL – Digital Media User Types of the Older Unemployed 40

    2.2 Activating Older Unemployed Individuals: A Case Study of Online Job Search Peer Groups 60

    2.3 Leveraging the Power of Peer Groups for Refugee Integration: A Randomized Field Experiment Comparing Online and Offline Peer Groups 71

3. Research on Societal Challenges Posed by Digital Technologies 95

    3.1 Evaluating Explainable Artificial Intelligence – What Users Really Appreciate 95

    3.2 Fostering Human Agency: A Process for the Design of User-Centric XAI Systems 115

    3.3 Explainable Artificial Intelligence in Information Systems – A Review of the Status Quo and Future Research Directions 134

List of Figures

Figure 1	Overview of the dissertation’s research subjects.	3
Figure 2	Overview of the dissertation’s research topics.	7
Figure 3	Overview of the randomized controlled experiments in papers 2 and 3.	19
Figure 4	Experimental setup in papers 4 and 5.	20
Figure 5	Overview of the structure of the dissertation.	22

List of Tables

Table 1	Overview of this dissertation's research objectives and papers.	15-16
Table 2	Overview of this dissertation's research paradigms, research approaches, and data.	21-22

List of Abbreviations

AI	Artificial Intelligence
BSR	Behavioural Science Research
COVID-19	Coronavirus Disease 2019
DSR	Design Science Research
HICSS	Hawaii International Conference on Systems Sciences
IS	Information Systems
MISQ	Management Information Systems Quarterly
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MMD	Maximum Mean Discrepancy
RO	Research Objective
SDGs	Sustainable Development Goals
XAI	Explainable Artificial Intelligence

1. Introduction

This chapter motivates and describes the dissertation's research topics, followed by introducing the research objectives, the applied methodology, and the underlying research paradigms. This chapter concludes with an overview of the structure of the dissertation.

1.1 Motivation

Thomas John Watson assumed a "world market for about five computers" (Carr 2008); Steve Jobs assessed mobile devices "are OK if you're a reporter and trying to take notes on the run. But for the average person, they're really not that useful" (Sheff 1985). Exceeding expectations, digital technologies' "cumulative impacts have become so deep, wide-ranging and fast-changing as to herald the dawn of a new age" (United Nations 2019, p. 6). This "new age" (United Nations 2019, p. 6) is powered by digitization, defined as the "technical process of converting analog signals into a digital form, and ultimately into binary digits" (Legner et al. 2017, p. 1). In turn, this technological change fosters digitalization with "manifold sociotechnical phenomena and processes of adopting and using these technologies in broader individual, organizational, and societal contexts" (Legner et al. 2017, p. 1), with most of humanity being part of the online community of almost 5 billion internet users (International Telecommunication Union 2021), empowered to "connect to anyone else, obtain and generate knowledge, or engage in commercial or social activity" (United Nations 2019, p. 6).

Humanity's reliance on digital technologies in this "new age" (United Nations 2019, p. 6) is showcased by their role during the "first pandemic in the global era of widespread mobile-device-supported social media, Big Data and AI" (Gaffield 2020, p. 1; World Health Organization 2020). Technology is a critical part of the response to the Coronavirus Disease 2019 (COVID-19) pandemic (cf. Lalmuanawma et al. 2020; Whitelaw et al. 2020): The ubiquity of mobile phones is leveraged by gathering GPS locations from individual phones as well as telecom geolocation data to identify changes in population-level mobility, assess the risk level per region, trace outbreaks, and inform forecasting (Grantz et al. 2020). Additionally, records of proximal interactions between Bluetooth-enabled devices allow monitoring changes in regional pairwise contacts, facilitating individual contact tracing and quarantine (Ferretti et al. 2020; Grantz et al. 2020; Dar et al. 2020). The QR code saved on a mobile device serves to certify the COVID-19 health status (Whitelaw et al. 2020). Web-based services such as the Johns Hopkins "COVID-19 Dashboard" distribute information about indicators, such as incidence or case-fatality ratio (Johns Hopkins University 2022). Machine learning (ML) supports COVID-19 diagnoses (Lalmuanawma et al. 2020), e.g., gradient-boosting models to predict a diagnosis (Zoabi et al. 2021), deep convolutional neural networks to classify radiology images (Li et al. 2020; Ardakani et al. 2020), support vector machines to predict severe symptoms (Sun et al. 2020) and mortality rates (Yan et al. 2020) in COVID-19 patients. Further, ML provides forecasts (Lalmuanawma et al. 2020), e.g., deep learning to predict cases and hospitalization rates (Zeroual et al. 2020). ML is even employed to augment treatment discovery (Lalmuanawma et al. 2020; Ekins et al. 2020), e.g., a deep learning model identifies drugs that could act on COVID-19 proteins (Beck et al. 2020). In addition to these applications, societies' digital connectedness serves as an infrastructure to enable physical distancing (Gaffield 2020).

Indeed, the fight against COVID-19 showcases humanity's reliance on digital technologies as "the most powerful new tool we have for solving the world's major challenges" (Sachs et al. 2016, p. 6). Research supports these expectations by establishing technology's potential to foster societal good across a wide range of challenges (e.g., Majchrzak et al. 2016). Thus, the United Nations expect this new "tool for social good" (Sachs et al. 2016, p. 6) to support the 17 Sustainable Development Goals (SDGs), outlining the global community's shared commitments for a better future (United Nations General Assembly 2015).

In contrast to these high expectations, COVID-19 also brought to light the perils that the "new age" entails (United Nations 2019, p. 6): Hospitals relied on ML tools that were not fit for clinical use (Heaven 2021; Borzyskowski et al. 2021). Analyzing 232 prediction models in a living review approach that is updated based on new evidence, Wynants and colleagues found that a vast majority of models aiming to diagnose and predict the course of COVID-19 in patients with suspected infections had insufficient quality, with only two models identified as promising (2020). Another review supports this assessment, finding no model fit for clinical translation after analyzing 415 studies focused on detection and prognosis based on chest radiographs and computed tomography (Roberts et al. 2021). Derek Driggs, a co-author of the latter review, concludes in an interview with the Massachusetts Institute of Technology (MIT) Technology Review that "this pandemic was a big test for AI and medicine. […] But I don't think we passed that test" (Heaven 2021).

ML's inherent limitations become apparent when studying the root causes of these failures (Liao 2020): Any model is only as good as the underlying training data, and an algorithm might result in models that fail to identify signals or falsely recognize a non-existing pattern, picking up on spurious correlations (Liao 2020). This is illustrated by examples such as a model aiming to diagnose and predict the course of COVID-19 trained using a dataset as a control group that consists of patients aged between one and five (Kermany et al. 2018), with the model ultimately learning to identify children versus adults and failing to pick up on COVID-19 symptoms (Roberts et al. 2021). Another example is a model to identify hip fractures that picked up the physician's markers in x-ray images (Badgeley et al. 2019). Moreover, ML might reinforce an unfair treatment of protected classes captured in historical data (for an overview, see Mehrabi et al. 2021). A lack of diversity and representation of minority groups in datasets used to train ML biases the decision-makers it seeks to inform (Borzyskowski et al. 2021), e.g., an ML-based hiring tool employed and later discontinued by Amazon discriminated women, perturbating existing bias in hiring decisions (Dastin 2018). Overall, the opaque nature of many ML systems elevates these limitations, as it impedes scrutiny and hinders identifying such pitfalls. Deployed in high-risk environments, ML might lead to fatal outcomes, as showcased by a facial recognition system causing an innocent person's arrest (McGregor 2020).

Societal risks stemming from the application of digital technologies are not only caused by technical limitations. They can also arise due to human vulnerabilities (Liao 2020): During the COVID-19 pandemic, social media platforms like Twitter radically accelerated the spread of misinformation (Kouzy et al. 2020), health-related conspiracy theories (Allington et al. 2021), ultimately contributing to a decrease in the likelihood of compliance with public health guidance (Roozenbeek et al. 2020). Generative adversarial networks allow bending the line between synthetic and real content (Liao 2020; Verdoliva 2020). While this allows for, e.g., new creative expression (Xue 2021) and the creation of additional data to balance a training dataset (Zhang et al. 2020), it also poses severe risks by fueling disinformation campaigns and leading to electoral distortion (Liao 2020). As society relies on digital connectedness to facilitate physical distancing,

the digital divide amplifies socioeconomic inequalities (Whitelaw et al. 2020; Gaffield 2020). Indeed, the Global Risk Report published by the World Economic Forum highlights digital power concentration, digital inequality, and cyberattacks as threats to humanity (World Economic Forum 2021).

The promise and perils of digital technologies are two sides of the same coin, as "opportunities created by the application of digital technologies are paralleled by stark abuses and unintended consequences. Digital dividends co-exist with digital divides" (United Nations 2019, p. 4). With society standing at the "dawn of a new age" (United Nations 2019, p. 6), digital technologies are both tools for societal good (Sachs et al. 2016) and risks to humanity (World Economic Forum 2021). Thus, the promise and perils of digital technologies are crucial to society and demand research attention. "Information systems (IS) as a field of academic research and business practice has long considered the importance of ethical considerations, including questions of what counts as right and wrong, good or bad, moral or immoral. […] Such questions touch on the design and use of computing artefacts in organizations in many different ways." (Stahl et al. 2014, p. 810). "Responsible Research and Innovation" in IS is linked to computer ethics, while focusing less on providing philosophical-theoretical contributions but rather on investigating practical implications of digital technologies (Stahl et al. 2014). While "Responsible Research and Innovation" initially focused on "preventing harm arising from research activities" (Stahl et al. 2014, p. 814), it has widened its goals toward responding to "grand challenges" that are "global or cover large parts of humanity" (Stahl et al. 2014, p. 814). In line with these dual goals, Responsible Technology focuses on "questions around how technologies can be conceptualized, designed, deployed or used in ways that are conducive or detrimental to human happiness" (Jirotka and Stahl 2020, p. 1). Against this background, this dissertation aims to foster research on both capitalizing on the power of digital technologies to alleviate societal challenges and addressing the challenges it poses to society (cf. figure 1).

Figure 1: Overview of the dissertation's research subjects.

This dissertation addresses selected aspects of how digital technologies can alleviate societal challenges (*Subject A*), as they are increasingly perceived to be tools for societal good (Sachs et

al. 2016; United Nations 2019), contributing to "promote human flourishing" (Bynum 2006, p. 157). Seeking to understand if digital technologies can live up to these expectations, this dissertation focuses on the question of whether and how they can be employed to assist marginalized communities (AbuJarour et al. 2019), i.e., "populations outside the mainstream society" (Cheraghi-Sohi et al. 2020, p. 1). Indeed, the potential to leverage digital technologies to support marginalized communities seems to exist (AbuJarour et al. 2019): Digital technologies in general, and especially the rise of mobile technologies, allow ubiquitous access to a diverse range of services, facilitate low-cost deployment, and enable rapid uptake at a global scale (Sachs et al. 2016). Successful examples include the person-to-person money transfer system M-PESA in Kenya, which lifted 194,000 households out of poverty (Suri and Jack 2016), and eHealth kiosks that disseminate medical information to fight infant mortality (Venkatesh et al. 2016). What is more, research postulates that digital technologies can connect geographically dispersed individuals and, thus, foster communication and collaboration. One example is a network of people "who have come together for mutual assistance in satisfying a common need, overcoming a handicap or bringing about desired social and/or personal change" (Katz and Bender 1976, p. 278). Such online peer groups successfully support individuals across various challenges ranging from improving social participation in the elderly (Goswami et al. 2010) to assisting women with postpartum depression (Prevatt et al. 2018).

Despite this potential, research (AbuJarour et al. 2019; Majchrzak et al. 2016) and policy (United Nations 2019) are only starting to capitalize on the power of digital technologies for marginalized groups and call to empirically validate technology's impact in this context (AbuJarour et al. 2019; Majchrzak et al. 2016). Responding to this call, this dissertation aims to expand research on the power of digital technologies to alleviate societal challenges by observing, analyzing, and designing solutions to exploit their benefits in marginalized groups. This dissertation's guiding question for Subject 1 is whether and how digital technologies can assist marginalized groups, focusing on older, unemployed individuals and refugees.

Unemployment in older individuals (*Topic 1*) is a severe societal challenge. A rapidly growing portion of the worldwide population becomes financially dependent, threatening public finances and economic growth (OECD 2019; United Nations 2020). Research shows that older individuals experience higher financial and psychological losses related to unemployment (Jebb et al. 2020; Klehe et al. 2012). Further, job loss is often a prelude to long-term unemployment and an early labor market exit (OECD 2019; United Nations 2020), as older individuals have lower chances of returning to work (Tisch 2015; Vansteenkiste et al. 2015).

Although digital solutions might help to provide broad access to job-search assistance, facilitating low-cost deployment at scale with fewer time and space constraints for participants (Belot et al. 2019; McQuaid et al. 2004), most active labor market programs are still delivered in person (Biewen et al. 2007; Card et al. 2018; OECD 2019). While initial research testifies to the benefits of digital technologies for finding employment in young and middle-aged individuals (Felgenhauer et al. 2019; Garg and Telang 2012, 2018; Klier et al. 2019; Kuhn and Mansour 2014), only sparse information exists about the effectiveness in the older, unemployed population. Therefore, the potential of digital technologies to improve the re-employment chances in this group remains untapped (Liu et al. 2014; OECD 2019).

A study investigating the effectiveness of digital interventions on re-employment indicators across age groups found no improvement in those above 50 (Briscese et al. 2020). The authors of this study suggest a reduced ability to navigate digital tools as an explanation for this observation

(Briscese et al. 2020). An alternative answer grounded in research on offline job search interventions is that older job seekers benefit most from interventions explicitly targeted toward that age group (Liu et al. 2014; Boockmann and Brändle 2019). The potential for societal impact of digital technologies on the older unemployed is supported by research indicating that digital technologies might help older individuals, e.g., improving cognitive performance (Ordonez et al. 2011), strengthening empowerment (Hill et al. 2015), fostering connectedness (Lüders and Brandtzæg 2017), and supporting mental well-being (Cotten et al. 2012). In addition, social media engagement promoted well-being in unemployed participants across age groups, while the positive effect was especially noticeable in older participants (Suphan et al. 2012). Considering these seemingly contradictory findings, this dissertation sheds light on the potential of digital technologies to help fight unemployment in older individuals.

The integration of refugees (*Topic 2*) poses a tremendous challenge for both host country communities and the individual who "owing to well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country" (United Nations General Assembly 1951, p. 14). In mid-2021, the number of people forcibly displaced due to prosecution, conflict, or generalized violence, reached an unprecedented peak of 84 million (United Nations High Commissioner for Refugees 2021), with an additional 4.8 million people fleeing Ukraine since February 2022 (United Nations High Commissioner for Refugees 2022). Local integration is one durable solution to address the consequences of this displacement (United Nations High Commissioner for Refugees 2020). Still, especially given the extraordinary number of individuals, it often poses tremendous challenges for refugees and their host countries. Risks include the potential for refugees' long-term financial dependency on their host countries, isolation of individuals, marginalization as a group, and increasing political radicalization (United Nations High Commissioner for Refugees 2013).

Research in IS is criticized for studying the impacts of digital technologies in mainstream communities, gaining insights that are only partially transferrable to refugees (AbuJarour et al. 2019). Still, initial research shows digital technologies' potential to support refugees from pre-departure to integration (Benton and Glennie 2016). The importance of digital technologies during flight is illustrated by refugees reporting that "internet is the same like food" (Kutscher and Kreß 2016, p. 1). Mobile phones serve many purposes, including connectivity and navigating toward better routes (Eide 2020; United Nations High Commissioner for Refugees 2016), while smartphones provide companionship, diversion, and facilitate organization during flight (Alencar et al. 2019; United Nations High Commissioner for Refugees 2016). Further, research shows that virtual networks and mobile phones expand higher education opportunities for women in refugee camps (Dahya and Dryden-Peterson 2017). Digital technologies also support refugees upon arrival in the host country: Digital applications aggregate local information about the host community (Schreieck et al. 2017a, 2017b), promote local events, and help refugees learn a new language (Ngan et al. 2016). What is more, they allow refugees to help each other by providing a platform for, e.g., health-related (Benton and Glennie 2016) or overarching issues (Schäfer-Siebert and Verhalen 2021).

Given these successes in helping refugees from pre-departure to arrival, there is a call for research leveraging digital technologies to support the phase of long-term integration into the host country community (AbuJarour et al. 2019). Initial research supports this quest, as Siddiquee and Kagan identify that an intervention to teach internet skills fostered empowerment and participation

in female refugees in the United Kingdom (2006). Other findings show that digital technologies can enhance refugees' well-being, empower them toward better situational control, and increase their societal participation (Díaz and Doolin 2016). Bacishoga and Johnston show that mobile phones positively affect social, cultural, and economic participation (2013), and online networks facilitate refugees' social connections and learning (Alencar 2018). Against this background, this dissertation investigates the social impact of technology-based interventions and answers the call for empirical investigations of digital technology's impact on long-term integration (AbuJarour et al. 2019).

With *Subject B*, this dissertation aims to contribute to research on societal challenges posed by digital technologies. It focuses on Artificial Intelligence (AI), expected to "have a more profound impact on humanity than fire, electricity and the internet" (Knowles 2021, p. 1), while posing risks of "increased gender and ethnic bias, significant threats to privacy, dignity and agency, dangers of mass surveillance, and increased use of unreliable Artificial Intelligence technologies in law enforcement, to name a few" (United Nations 2021, p. 1).

"AI" was coined by John McCarthy as "making a machine behave in ways that would be called intelligence if a human were so behaving" in the foundational Dartmouth Summer Research Project on AI (McCarthy et al. 2006, p. 11). The research field of AI can be characterized as "the study and construction of agents that do the right thing" (Russell and Norvig 2021, p. 22). To date, the driving technology fostering AI's rapid progress is ML, i.e., the capability to learn from examples based on the following process: "a computer observes some data, builds a model based on the data, and uses the model as both a hypothesis about the world and a piece of software that can solve problems" (Russell and Norvig 2021, p. 669). Both the issues that arise due to human vulnerabilities and vulnerabilities in ML itself (Liao 2020) illustrate that "without AI systems [..] being demonstrably worthy of trust, unwanted consequences may ensue" (HLEG-AI 2019, p. 4). In response, various stakeholders, including policy, society, and businesses, create guidelines and approaches to pave the way toward "responsible AI" that aim at building fair, accountable, and explainable systems (Arrieta et al. 2020, p. 1).

One prominent example of regulation is the requirement of providing "meaningful information about the logic involved [..] for the data subject" (European Parliament and the Council of the European Union, General Data Protection Regulation, Article 13 – 15, 2016). Another example of upcoming regulation is the European Union's AI Act: The Act groups AI systems into three categories distinguished by the perceived risk to the public. It requests not to deploy systems that pose an 'unacceptable risk' to society. Systems of the second risk category are called 'high risk' systems and have to comply with several requirements, including the provision of transparency and human oversight (European Commission 2021). Both these regulations address the opaque nature of many ML systems. One example relevant across various applications from visual object recognition to natural language processing is neural networks consisting of artificial neuron layers representing complex nonlinear functions (Russel and Norvig 2021; Goodfellow et al. 2016). These models are opaque as they cannot be understood by looking at the parameters (Molnar 2022). Therefore, the reasons driving their recommendations and predictions appear unfathomable to users (Guidotti et al. 2020).

"In order to be beneficial to individuals and society, the proliferation of ML in everyday life requires that users are able to comprehend and interact with ML systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system" (HLEG-AI 2019,

p. 16). Yet, when dealing with opaque systems, users lack sufficient information to reflect critically (Rader and Gray 2015) and make an informed decision about the ML system's recommendations (Guidotti et al. 2020; Mittelstadt et al. 2019). In turn, they cannot contest, appropriately trust, and manage their ML partners (Wachter et al. 2018; Arrieta et al. 2020). Additionally, this creates a barrier to ML adoption (Arrieta et al. 2020), as users have the choice of either blindly following an ML recommendation or merely distrusting and not using it (Rader and Gray 2015). Moreover, ML engineers and data scientists lack tooling for debugging models and identifying issues such as spurious correlations or undesirable bias (Bhatt et al. 2020). In light of these challenges, this dissertation contributes to the emerging research field of Explainable AI (XAI) that seeks to drive responsible ML adoption by providing explanations accompanying the ML system's outputs.

1.2 Research Objectives

The objective of this dissertation is to contribute to research on digital technologies to alleviate societal challenges *(Subject A)* and respond to societal challenges posed by digital technologies *(Subject B)*. Subject A comprises two topics (Topic I: Unemployment in Older Individuals; Topic II: Integration of Refugees), while subject B focuses on one topic (Topic III: Explainable AI), as illustrated in figure 2. In the following, the research objectives are motivated and introduced.

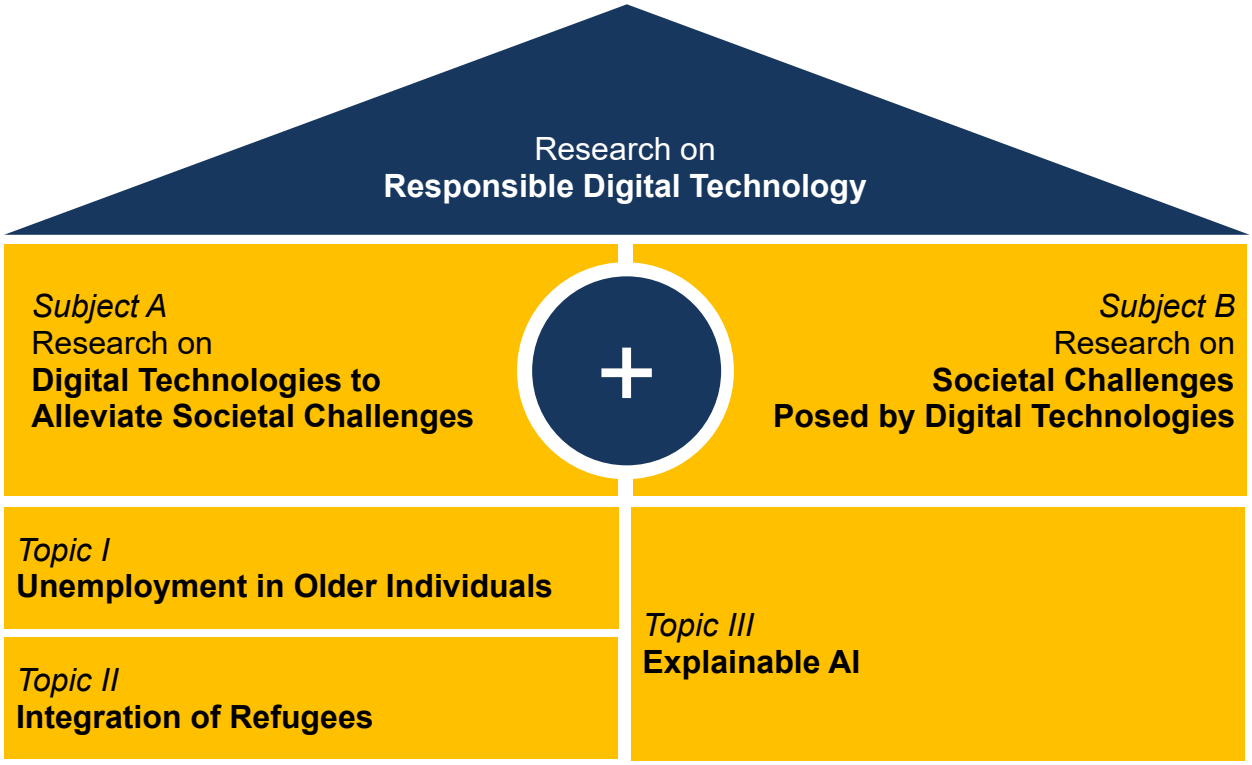


Figure 2: Overview of the dissertation's research topics.

Subject A aims to understand how digital technologies contribute to alleviating societal challenges. While digital technologies are perceived as a tool to foster human welfare (Sachs et al. 2016; United Nations 2019) research and policy are only starting to capitalize on the power of digital technologies for marginalized communities (AbuJarour et al. 2019). In investigating the

potential benefits in this context, this dissertation focuses on two topics: unemployment in older individuals and integration of refugees.

Topic I: Unemployment in Older Individuals

"Digital dividends co-exist with digital divides" (United Nations 2019, p. 4). While digital technologies might support older unemployed individuals, increasing age and unemployment also raise the risk of entering a "digital underclass" (Helsper 2011, p. 1; Helsper and Reisdorf 2017).

Older individuals are less likely to use digital technologies (Hunsaker and Hargittai 2018; König et al. 2018), e.g., as shown by their low adoption of eGovernment services (Niehaves and Becker 2008), with a similarly low digital technology uptake found in unemployed individuals (Helsper and Reisdorf 2017). Yet, initial research suggests that age and socioeconomic status have no explanatory power for a reluctance to use digital technologies when controlled for attitude and experience (Siren and Knudsen 2017). These findings are supported by research highlighting that causes for the lack of digital technology usage in older individuals include an overall negative attitude toward digital technologies, perceived lack of utility, and concerns around data privacy (Olphert et al. 2005; Morris et al. 2007; Lüders and Brandtzæg 2017). Further lack of internet access (Morris et al. 2007; König et al. 2018), insufficient digital skills (Olphert et al. 2005; Lüders and Brandtzæg 2017), or the perception of high costs (Lee et al. 2011), are identified as barriers to usage in older individuals.

In light of this research investigating the causes driving low digital technology usage, one should stop treating older, unemployed individuals as one homogenous group and distinguish between different user types and characteristics. Thus, as a first step in understanding whether digital technologies can help older individuals facing unemployment, this dissertation extracts user typologies (Barnes et al. 2007; Brandtzæg 2010). A canonical example of user typologies that served as a foundation for later research is Brandtzæg's "new media" typology addressing the usage of "television, computers, internet, different game consoles, mobile phones" (2010, p. 943). Users are grouped into clusters ranging from Non- to Advanced Users. The so-called Sporadics and Lurkers are between the extremes, the former characterized by low frequency and variety of usage and the latter by passive consumption (Brandtzæg 2010). In addition to these types, there are four types with a medium frequency and variety of usage: Debators actively contribute to blogs and social networking sites, whereas Socializers use social media to nurture relationships. Instrumental Users focus on gathering utility and information, while Entertainment Users seek fun and distraction (Brandtzæg 2010). Literature offers a wide variety of studies identifying user types of specific technologies, such as social media (e.g., Akar et al. 2019; Kim 2018), and shedding light on particular user groups, such as children (Zawacki-Richter et al. 2015) and the unemployed (Feuls et al. 2016).

To the best of my knowledge, no digital technology user typology of older, unemployed individuals was developed prior to this research endeavor. Thus, to address the challenges of better understanding the digital media preferences and characteristics of older, unemployed individuals, this dissertation addresses the following research objective (RO):

- *RO1: Explore the different digital technology user types of the older unemployed.*

As a second step, this dissertation aims to shed light on whether digital technologies can assist the older, unemployed population. While initial research shows the power of digital technologies to support job search in young and middle-aged individuals (Felgenhauer et al. 2019; Garg and

Telang 2012, 2018; Klier et al. 2019; Kuhn and Mansour 2014), most services to support job search in the older unemployed are delivered in person and research on the effectiveness of digital technologies for this group remains sparse (Biewen et al. 2007; Card et al. 2018; OECD 2019).

To tap into the potential of digital technologies to support the reemployment chances of the older unemployed, this dissertation investigates the introduction of a digital labor market intervention in a randomized field study at the Federal Employment Agency in Germany between February 2019 and March 2020. Older, unemployed participants got access to an online peer group that provided a forum for discussion facilitated by digital technology. This brought together individuals who offered each other assistance in dealing with unemployment at an older age (Katz and Bender 1976). This digital intervention was previously found to support participants across various challenges, from assisting women with postpartum depression (Prevatt et al. 2018) to helping youths find a job (Klier et al. 2019). Research identifies that peers in these (online) peer groups provide five types of social support: informational support, emotional support, esteem support, network support, and tangible assistance (Cutrona and Suhr 1992).

Online peer groups have the potential to support job search, as positive effects observed in such groups reflect indicators of elevated reemployment (Liu et al. 2014; Wanberg et al. 2002; McQuaid 2006): First, online peer groups foster *knowledge gain*, e.g., helping parents to learn about parenting (Niela-Vilén et al. 2014). Knowledge gain is essential for reemployment as improved job search skills, e.g., interviewing skills, elevate job search outcomes (Liu et al. 2014; Wanberg et al. 2002; McQuaid 2006). Improving job search skills is particularly important in older individuals, as they are more likely to lack the ability to navigate digital job search tools (Tisch 2015; OECD 2019) and have worse job search skills overall (Tisch 2015; Liu et al. 2014). Second, peer groups can induce positive *behavior change*, e.g., strengthening career search intensity in youths (Klier et al. 2019). Devoting more time and effort to job search, e.g., submitting a higher number of applications is related to better reemployment outcomes (Wanberg et al. 2002; McQuaid 2006; Schmidt 2007). Fostering job search behaviors is vital for the older unemployed, as job search intensity declines with age and partly accounts for the lower reemployment rates observed in older unemployed (Vansteenkiste et al. 2015; Rife and Belcher 1994). Third, peer groups elevate *self-efficacy*, i.e., the "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (Bandura 1997, p. 3). When participating in a peer group, individuals with stigmatized chronic diseases improved in self-care behaviors (Wang et al. 2017). Self-efficacy increases the chances of re-entering the labor market (Fugate et al. 2004) and declines with age (Maurer 2001). Finally, the positive effects of peer groups might also support older individuals to better cope with job loss. Peer groups intensify *social connectedness* and general *well-being*, e.g., improving social participation in the elderly (Goswami et al. 2010) and assisting women with postpartum depression (Prevatt et al. 2018).

While the findings presented above suggest that online peer groups have the potential for supporting job search in the older unemployed, so far, it has not been empirically investigated. This dissertation focuses on addressing the research gap of exploiting the potential of digital technologies, specifically online peer groups, to assist the job search of the older unemployed by addressing the following research objective:

- *RO2: Investigate if digital technologies improve reemployment chances in older, unemployed individuals.*

Topic II: Integration of Refugees

Research suggests that digital technologies can support refugees (e.g., Siddiquee and Kagan 2006; Bacishoga and Johnston 2013; Dahya and Dryden-Peterson 2017), allowing ubiquitous access to a wide range of services and connection to geographically dispersed family and friends (AbuJarour et al. 2019). Positive outcomes of digital interventions in this context include enhanced well-being, better situational control, increased societal participation (Díaz and Doolin 2016), improved social connections, and elevated learning accomplishments (Alencar 2018). While the successes of digital technologies were established along several stages, from pre-departure to arrival in the host country (e.g., Eide 2020; Alencar et al. 2019; Dahya and Dryden-Peterson 2017; Schreieck et al. 2017a, 2017b; Ngan et al. 2016; Benton and Glennie 2016; Schäfer-Siebert and Verhalen 2021) there is a call for research to also focus on supporting the long-term integration (AbuJarour et al. 2019). Providing tools to manage long-term integration better is a crucial building block to solving displacement, as returning to their home countries is not possible for many refugees (United Nations High Commissioner for Refugees 2020).

Integration is defined as a concept based on "adaptation" and "welcome," consisting of economic, legal, and social-cultural dimensions (United Nations High Commissioner for Refugees 2013). Based on that definition, research proposes several frameworks to capture aspects of successful integration (e.g., AbuJarour et al. 2018; Harder et al. 2018). One of these frameworks that is applied widely in policy and research (e.g., Hynie et al. 2016) was proposed by Ager and Strang (2008). The authors identify four core integration themes, that encompass several domains (Ager and Strang 2008): *Understanding of citizenship and according rights* provides the foundation, on which *Language and Cultural Knowledge,* as well as *Safety and Stability,* serve as enablers of successful integration (Ager and Strang 2008). The penultimate layer focuses on a threefold view of social connections, i.e., to the host country community, its public structure, and the home culture (Ager and Strang 2008). Finally, *Employment, Housing, Education,* and *Health* are both drivers and outcomes of successful integration (Ager and Strang 2008).

Similar to the case of the older, unemployed individuals discussed above, online peer groups seem to be a promising intervention in this context, as the positive effects observed in other contexts relate to indicators supporting long-term integration (Ager and Strang 2008). *Knowledge gain* (e.g., Niela-Vilén et al. 2014) is an essential aspect of integration: Refugees might have to study a new language upon arrival or learn about the host country's culture and social functioning (cf. Ager and Strang 2008). In addition, refugees frequently must find new employment and get access to vocational training, housing, or healthcare (cf. Ager and Strang 2008), the positive *behavior change* induced by peer groups (e.g., Klier et al. 2019) might be crucial in tackling these challenges. In addition to that positive change in behavior, an increase in *self-efficacy* (e.g., Barak et al. 2008) might support refugees through setbacks and difficulties, elevating the "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (Bandura 1997, p. 3). Intensifying *social connectedness* (e.g., Goswami et al. 2010; Felgenhauer et al. 2019) might support the triple layer of social connections required for successful integration, namely maintaining ties to the home culture and building new relationships with the host community and institutions (cf. Ager and Strang 2008). Finally, increasing general *well-being* (e.g., Prevatt et al. 2018) might support refugees in dealing with traumatic experiences and regaining a sense of safety and stability (cf. Ager and Strang 2008).

In summary, the effects observed in peer groups might support the indicators of successful integration outlined by Agar und Strang (2008). Yet, research neglected empirically demonstrating

the benefits of using (online) peer groups to support refugees' long-term integration. Against that background, this dissertation pursues the following research objective:

- *RO3: Develop and evaluate a novel digital artifact to enhance refugee integration.*

The call for digital interventions to support marginalized communities (AbuJarour et al. 2019) is reinforced by research demonstrating the potential of technologies to empower refugees (AbuJarour et al. 2021; Díaz and Doolin 2016). Indeed, studies identify a strong proliferation of mobile phones and smartphones in refugees (Betts et al. 2017) used both during the flight from the home country (e.g., Dekker et al. 2018; Alencar et al. 2019) and upon arrival in the host county (e.g., Kaufmann 2018).

When designing services to support refugees, digital technologies allow for several advantageous design choices: Services powered by digital technologies require no co-presence and can support their users independent of location and time (e.g., Coulson 2013). In the context of refugee integration, this characteristic might provide instant assistance in their daily lives, e.g., when dealing with host country bureaucracy or connecting with friends and family at home. Further, digital services can power asynchronous and written interaction via messages, which allows users more time to respond (Andresen 2009), enables quick review of older exchanges (Bender et al. 2013), and lowers communication barriers (Braithwaite et al. 1999), which is crucial given that users interact in a foreign language and discuss the sensitive topic of seeking refuge in a foreign country. Also, non-co-presence allows for anonymous interaction, which supports communication on sensitive issues, e.g., discussing medical issues (Bender et al. 2013). In addition to that, after successful development, digital services allow deployment and vast distribution at a lower marginal cost than in-person services and facilitate continuous updates (Sachs et al. 2016).

Still, services based on digital technologies might also entail numerous disadvantages: One example is that non-copresence might make non-verbal expressions, which are especially important when communicating in a foreign language, more challenging (Kiesler et al. 1985) and hinder feelings of closeness between participants (Sannomiya and Kawaguchi 1999), hampering the establishment of social connections needed for successful integration (Ager and Strang 2008). Thus, when designing a digital solution, careful comparison with the face-to-face equivalent is required in order to explore the respective advantages and limitations (e.g., Rupert et al. 2017; Duryea et al. 2021). Indeed, research calls for shedding light on the relative importance of online characteristics in interventions such as peer groups (Klier et al. 2019).

This dissertation addresses the lack of research comparatively assessing the effectiveness of an online and in-person implementation of peer groups in the context of refugee integration:

- *RO4: Understand the impacts of digital technologies in the context of refugee integration by comparatively assessing online and offline interventions' effectiveness.*

Subject B comprises of one topic: Explainable AI.

Topic III: Explainable AI

"Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand" (Arrieta et al. 2020, p. 6). Thus, explainable systems aim to produce a rationale that is comprehensible to both data scientists who understand the underlying algorithmic structure and to users without a technical background (Chakraborty et

al. 2017; Abdul et al. 2018). While some ML models are explainable by design, e.g., linear regression or decision tree, others require applying methods to achieve explainability (Arrieta et al. 2020; Molnar 2022). Such post-hoc explainability involves the application of a method to an ML model after training (Doran et al. 2018; Guidotti et al. 2020).

To achieve post-hoc explainability, one can use model-agnostic or model-specific methods (Arrieta et al. 2020; Molnar 2022). Model-agnostic XAI methods provide explanations applicable to a range of ML algorithms. One prominent group of explanation methods are Shapley Additive Explanations, based on a game-theory method for credit attribution called Shapley Values (Shapley 1953). These explanations assign a value per input feature indicating its importance for an individual prediction (Lundberg and Lee 2017). On the other hand, model-specific methods focus on explaining a particular group of ML algorithms, e.g., methods tailored to explain deep neural networks (Montavon et al. 2018).

A second axis to group explainability methods is to distinguish between local and global methods. Former aim at explaining an individual prediction, such as the Shapley Additive Explanations discussed above (Lundberg and Lee 2017). Another example of local explanations are contrastive explanations that highlight why a specific outcome was generated, e.g., credit line rejection, instead of the counterfactual situation, e.g., credit line approval (e.g., Guidotti et al. 2020; van der Waa et al. 2018; Wachter et al. 2018). In contrast to that approach, global explanations aim at explaining the behavior of an ML model, not just an individual prediction, e.g., Maximum Mean Discrepancy (MMD)-critic (Kim et al. 2016): This framework identifies prototypes, i.e., instances at the center of a data distribution, representative of a data cluster, and criticisms, i.e., data points that are not well represented by these prototypes and do not fit the model (Kim et al. 2016).

As evident by the variety of XAI methods highlighted above, the call to generate explanation methods for ML systems has attracted considerable research attention (Arrieta et al. 2020; Guidotti et al. 2020). Still, research to date is criticized for a situation best described as "inmates running the asylum" (Miller et al. 2017), with researchers constructing explanations they appreciate rather than explanations that generate value for their users (Kirsch 2018; Miller 2019; Mittelstadt et al. 2019): "Researchers in the ML and AI communities are working on making their algorithms explainable, their focus is not on usable, practical and effective transparency that works for and benefits people" (Abdul et al. 2018, p. 10).

Against this background, this dissertation focuses on putting the user at the center of research attention. The aim is to design explanations that users appreciate. Social science research provides insights into how humans construct and perceive explanations (Miller 2019) and finds that an explanation's "loveliness" contributes to its "likeliness" (Lipton 2000). Thus, specific explanation characteristics elevate a user's appreciation. First, recipients prefer short explanations in most settings, i.e., with a lower "number of causes invoked in an explanation" (Lombrozo 2007, p. 232). Yet, specific settings such as explaining a scientific phenomenon yield a preference for longer explanations (Weisberg et al. 2015). Second, explanations are preferred to be consistent and related to a recipient's existing beliefs, i.e., coherent (Thagard 1989; Lombrozo 2012). Third, recipients favor general explanations, i.e., applicable to a larger number of observations and including root causes (Thagard 1989; Lombrozo 2012). Finally, relevance (Hilton and Erb 1996; McClure 2002) contributes to an explanation's "loveliness", as recipients favor causes that are of high proximity to the outcome (Miller and Gunasegaram 1990), neither surprising (Hilton and Slugoski 1986) nor abnormal (Kahneman and Tversky 1981; McCloy and Byrne 2000).

Still, these characteristics can only provide a starting point in constructing explanations for ML systems, as humans are known to act differently when their counterpart is not human (e.g., Rzepka and Berger 2018). Yet, there is a lack of empirical investigation into characteristics of explanations in the context of interaction with an AI system that users appreciate (Miller et al. 2017; Kirsch 2018; Mittelstadt et al. 2019). Thus, to build explanations that serve users, this dissertation conducts a human-based study to shed light on the following research objective:

- *RO5: Investigate characteristics of explanations generated by XAI systems that users appreciate.*

When evaluating explanation methods, research proposes three evaluation scenarios (Doshi-Velez and Kim 2018): Functionally-grounded evaluation is a theoretical investigation into explanation methods that does not require the involvement of human subjects, e.g., investigating problems with Shapley-value-based explanation methods by applying the methods to a variety of ML modes and datasets (Kumar et al. 2020). On the other hand, both application-grounded and human-based evaluations require an experiment with human subjects. They differ in that application-grounded evaluation involves testing with the anticipated end-users in the intended application setting, e.g., evaluating methods to generate post-hoc explanations in a fraud detection task with fraud analysts (Jesus et al. 2021). Human-grounded evaluation, on the other hand, is performed on proxy tasks. In addition, it does not require the anticipated end-user to participate in a study, e.g., fraud analysts, but allows for evaluation with, e.g., participants recruited via Amazon's "Mechanical Turk" (Wang et al. 2016).

Only 5% of XAI research involves evaluating XAI methods (Adadi and Berrada 2018). Taking a closer look at the assessments performed, most involved functionally-grounded evaluation (Adadi and Berrada 2018), i.e., without testing with the arguably most important stakeholder group, namely users. To change that, research calls for more user-centered evaluations of XAI methods (Kirsch 2018; Miller 2019; Mittelstadt et al. 2019). This dissertation aims to address this quest by both performing a human-grounded evaluation of XAI methods and deriving a generic study design from that effort, thus addressing the following research objective:

- *RO6: Provide a generic study design to evaluate explanations generated by XAI systems.*

The European Commission urges AI systems to ensure human agency, as "users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system." (HLEG-AI 2019, p. 16). Explanations can serve as one fundamental building block in empowering users to that end (e.g., Wachter et al. 2018). Still, to achieve that goal explanations need to be designed in a user-centric manner (Ribera and Lapedriza 2019), focusing on the needs and wants of the target user (Norman and Draper 1986), while to date XAI research focuses "not on usable, practical and effective transparency that works for and benefits people" (Abdul et al. 2018, p. 10).

User-centric design might help bridge this chasm, as it aims to enhance the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO 2019, sec. 3.13). This aim translates to the field of XAI as the extent to which a user can "reasonably self-assess or challenge the system" (HLEG-AI 2019, p. 16). Several approaches exist to support user-centric

design (cf. Still and Crane 2017), e.g., guidelines stressing the importance of empathy and the integration of feedback mechanisms (IDEO 2015; Google 2019). The generic ISO framework proposes several principles and activities to guide human-centric design (ISO 2019) and is widely applied in business and academics, e.g., serving to inform user-centric software development processes (Farinango et al. 2015). The six principles set forth by the ISO framework start with an in-depth understanding of the target user, underlying task, and deployment environment. Second and third, it requires user involvement and continuous evaluation throughout the design and the development phases of the product lifecycle. The process is set out to be iterative and aims at encompassing the entire user experience. Finally, there is a quest to build a multidisciplinary and diverse product team (ISO 2019).

While initial research has started gathering insights and criteria for user-centric XAI design (Doshi-Velez and Kim 2018), e.g., from social sciences (Miller 2019) and user studies (Wang et al. 2019), these contributions remain disintegrated, and users are still not systematically incorporated into the development of XAI systems (e.g., Miller et al. 2017; Mittelstadt et al. 2019). Research in data mining and data sciences employs processes to address the challenge of translating requests, such as criteria for user-centric design and the fragmented insights into user-centric XAI, into technical requirements (Martinez-Plumed et al. 2019; Marbán et al. 2009). This dissertation takes processes employed in data mining and data science as a starting point to build a systematic guide for the design of XAI systems. Thus, this dissertation addresses the following research objective:

- *RO7: Develop and evaluate a novel process that guides the design of XAI systems toward fostering human agency.*

The call to explain ML models has generated numerous contributions in computer science (Arrieta et al. 2020) and attracted several scholars in the research field of IS (Meske et al. 2020). The emergence of XAI-focused editorials showcases this, e.g., the editorial "Expl(AI)n It to Me – Explainable AI and Information Systems Research" in Business & Information Systems Engineering (Bauer et al. 2021). Other examples include calls for papers, e.g., "Special Issue on Explainable and Responsible Artificial Intelligence" in Electronic Markets (Meske et al. 2022), "Special Issue on Designing and Managing Human-AI Interactions" in Information Systems Frontiers (Abedin et al. 2020), "Special Issue on Interpretable Data Science For Decision Making" in Decision Support Systems (Coussement and Benoit 2021). Finally, tracks in IS research conferences emerge, such as the "Minitrack on Explainable Artificial Intelligence" at the Hawaii International Conference on System Sciences (HICSS) (Meske et al. 2021).

Given the calls for user-centric XAI (e.g., Miller 2019) and the policy focus on ML explainability in AI regulation (e.g., HLEG-AI 2019, p. 16), there is a need to study ML explainability at the intersection between "people, organizations, and technology" (Hevner et al. 2004). In their editorial, Bauer et al. (2021) emphasize that IS research is predestined for this challenge, supported by further calls for more research on AI explainability in IS (e.g., Meske et al. 2020).

Still, prior to this dissertation, no research existed that summarized existing XAI research in IS. A comprehensive literature review is required to create transparency on existing research (Webster and Watson 2002), identify respective outlets (Bandara et al. 2011), and outline a research agenda for XAI in IS. Thus, this dissertation aims to address the following research objective:

- *RO8: Assess the status quo and identify potential future XAI research directions in IS literature.*

All these research objectives are addressed in 6 papers (cf. table 1).

Table 1: Overview of this dissertation's research objectives and papers.

Topic	Research Objective	Publication
Unemployment in Older Individuals	RO1: Explore the different digital technology user types of the older unemployed.	#JOBLESS #OLDER #DIGITAL – Digital Media User Types of the Older Unemployed • *Authors: J. Klier, M. Klier, K. Schäfer-Siebert, I. Sigler* • *Published in Proceedings of the European Conference on Information Systems (VHB-Rank: B)*
	RO2: Investigate if digital technologies improve reemployment chances in older, unemployed individuals.	Activating Older Unemployed Individuals: A Case Study of Online Job Search Peer Groups • *Author: I. Sigler* • *Published in Proceedings of the Hawaii International Conference on System Sciences (VHB-Rank: C)*
Integration of Refugees	RO3: Develop and evaluate a novel digital artifact to enhance refugee integration. RO4: Understand the impacts of digital technologies in the context of refugee integration by comparatively assessing online and offline interventions' effectiveness.	Leveraging the Power of Peer Groups for Refugee Integration: A Randomized Field Experiment Comparing Online and Offline Peer Groups • *Authors: M. Förster, J. Klier, M. Klier, K. Schäfer-Siebert, I. Sigler* • *Published in Business & Information Systems Engineering (VHB-Rank: B)*
Explainable AI	RO5: Investigate characteristics of explanations generated by XAI systems that users appreciate. RO6: Provide a generic study design to evaluate explanations generated by XAI systems.	Evaluating Explainable Artificial Intelligence – What Users Really Appreciate ▪ *Authors: M. Förster, M. Klier, K. Kluge, I. Sigler* ▪ *Published in Proceedings of the European Conference on Information Systems (VHB-Rank: B)*

Explainable AI	RO7: Develop and evaluate a novel process that guides the design of XAI systems toward fostering human agency.	Fostering Human Agency: A Process for the Design of User-Centric XAI Systems ▪ *Authors: M. Förster, M. Klier, K. Kluge, I. Sigler* ▪ *Published in Proceedings of the International Conference on Information Systems (VHB-Rank: A)*
	RO8: Assess the status quo and identify potential future XAI research directions in IS literature.	Explainable Artificial Intelligence in Information Systems: A Review of the Status Quo and Future Research Directions ▪ *Authors: J. Brasse, H. Broder, M. Förster, M. Klier, I. Sigler* ▪ *Submitted to Electronic Markets status: "In Review" (VHB-Rank: B)*

1.3 Research Paradigms and Research Methods

The discipline of IS investigates the intersection between "people, organizations, and technology" (Hevner et al. 2004). It provides the research foundation for the dissertation at hand, as the study of responsible technology requires an investigation along this intersection. Research in IS builds on two predominant and complementary paradigms, Behavioral Science and Design Science (Hevner et al. 2004). Behavioral Science Research (BSR) is grounded in methodologies established in the natural sciences and aims to explain human and organizational phenomena related to digital technologies by developing or verifying theories (Hevner et al. 2004). On the other hand, Design Science Research (DSR) is based on the engineering disciplines and aims to create novel artifacts, solving real-world problems (Hevner et al. 2004). These artifacts aim to be "readily converted to a material existence" (Gregor and Hevner 2013, p. 341) and can constitute constructs, models, methods, and instantiations (March and Smith 1995).

While in his original paper, Hevner specified the real-world problems as "business needs" (2004, p. 80), researchers have expanded this scope in leading IS journals such as the Management Information Systems Quarterly (MISQ). Manifesting that "the boundaries have broadened, from inside the organization to society and everything in between" (Goes 2013, p. iii), IS aims at "the development of IT-based services, the management of IT resources, and the use, impact, and economics of IT with managerial, organizational, and societal implications" (Goes 2013, p. i). Thus, societal implications and artifacts focused on generating societal impact are also at the focus of research attention, manifesting in calls for papers, e.g., "MISQ Special Issue on ICT and Societal Challenges" (Majchrzak et al. 2014) or "MISQ Special Issue on Digital Technologies and Social Justice" (Aanestad et al. 2022), as well as research focused on "digitally enabled solutions to cope with the wicked problems arising out of digitalization" (Benbya et al. 2020, p. 1). Thus, this expanded scope of IS research encompasses the research areas of this dissertation, namely, capitalizing on the power of digital technologies to alleviate societal challenges and addressing the challenges these technologies pose to society.

BSR and DSR are "two sides of the same coin" (Hevner et al. 2004, p. 77), and research is conducted in an iterative cycle between deriving knowledge by developing and justifying theories (BSR), which in turn informs the design and evaluation of artifacts (DSR). In describing this cycle, Hevner suggests a complementary relationship between generating truth and utility, forming the core of IS research which aims to deliver research outcomes to address real-world problems (2007). This dissertation relies on the complimentary research cycle of BSR and DSR (Hevner et al. 2004), to address the real-world problems of the need to solve societal issues by applying digital technologies and resolve perils caused by these technologies. The choice of research paradigm and methods is informed by the research objective of each paper (Hevner et al. 2004; Wilde and Hess 2007) (cf. table 2 for an overview) and presented in the following.

Subject A: Digital Technologies to Alleviate Societal Challenges

Topic 1: Unemployment in Older Individuals

This dissertation aims to understand whether digital technology can effectively assist the job search of older, unemployed individuals. To investigate this question, first, we gain insights into the digital technology preferences and characteristics of older, unemployed individuals and explore their different user types. Second, based on that understanding, we explore whether digital technologies improve this target group's reemployment chances. Thus, paper 1 and paper 2 follow the BSR paradigm to observe the potential of digital technologies to support job search in older individuals.

Paper 1 uses a quantitative, cross-sectional analysis (Wilde and Hess 2007) and builds on data gathered in a questionnaire-based survey with 192 valid participants. We chose surveys as the methodology for our study to take a user-centered perspective on digital technology usage, attitudes, and perceived barriers in the context of job search, following standard practice in IS research (Urbach et al. 2009). In addition to that, this methodological approach is in line with prior research on user typologies: The literature review provided by Brandtzæg (2010) highlights that questionnaire-based surveys are the most commonly employed means to derive user typologies. The survey was distributed to unemployed individuals above 50 in 19 Employment Agencies in the third-largest German state, spanning both urban and rural areas. We derived user groups as an established approach to gain insights into a new target group (Brandtzæg 2010), then we applied two-step clustering, followed by descriptive analyses and statistical tests to characterize and compare the clusters.

Complementary to the quantitative cross-sectional analysis conducted in paper 1, paper 2 employs an "in-depth inquiry into a specific and complex phenomenon (the 'case'), set within its real-world context" (Yin 2013, p. 1). The German Federal Employment Agency is selected to serve as the case setting. It allowed gathering a unique data set from a randomized controlled field experiment introducing a digital labor market intervention for unemployed individuals above 50 across seven different Employment Agencies in Germany.

Conducting a randomized controlled field experiment is the "gold standard for assessing cause and effect" (Liu and Wyatt 2011, p. 1). It allows the researcher a high level of control for unknown confounders compared to other evaluation approaches (Liu and Wyatt 2011). This is especially important when observing indicators related to employability as, e.g., uneven distribution in prior work experience or level of education between the treatment and control groups would distort results (Liu et al. 2014).

The findings build on two data sets, the usage data, with 205 treatment group participants, and the pre-and post-survey data, including 119 valid questionnaires for the treatment and 118 valid questionnaires for the control group. Descriptive analyses of the usage data serve to understand the approach's adoption, followed by statistical analyses to test for significant effects of the intervention.

Topic 2: Integration of Refugees

This dissertation contributes to conducting "more empirically grounded studies" on the potential of digital technologies to assist the long-term integration of refugees (AbuJarour et al. 2019, p. 15). Paper 3 applies the DSR paradigm to design and evaluate a novel approach to the societal challenge of refugee integration (Hevner et al. 2004). The novel artifact is based on the peer group approach (Katz and Bender 1976) and designed in two variants: A realization centered around face-to-face meetings (offline realization) and another based on communication via a mobile messaging solution (online realization). A randomized controlled field experiment in cooperation with the Federal Employment Agency in Germany and the German Red Cross at a so-called "Integration Point" in Heidelberg demonstrates the method's utility, quality, and efficacy (Hevner et al. 2004).

The choice of research method is similar to paper 2: The field experiment aims to observe whether the artifact provides substantial value to integration, while controlling for potentially confounding factors (Liu and Wyatt 2011), which is critical when observing indicators related to refugee integration given the impact of, e.g., elevated language skills on integration success (Ager and Strang 2008). Additionally, paper 3 was designed to directly compare the differences in the effectiveness of an online and an offline realization, thus allowing to distill the unique contribution digital technologies have to support refugee integration.

The findings build on three datasets: demographic data, usage data, i.e., data on participation in the online and offline realization of the peer-group-based artifact, and survey data. First, usage data analyzes the approach’s adoption, while statistical tests manifest a significant difference between the online (n = 65) and offline peer group (n = 63). Second, survey data served to observe differences in post- and pre-treatment changes between the online treatment group (n = 54), the offline treatment group (n = 53), and the control group (n = 51), while statistical tests were applied to test for the significance of these changes.

See figure 3 for an overview of the randomized controlled experiments conducted to achieve the research objectives in subject A.

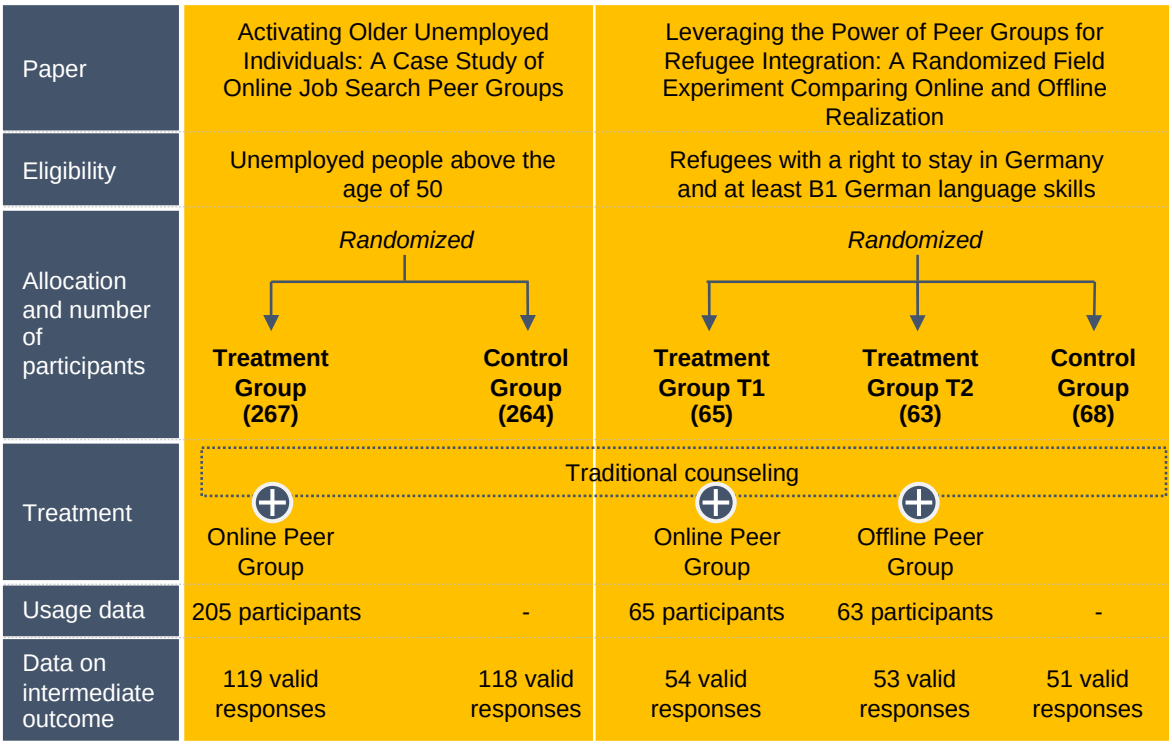


Figure 3: Overview of the randomized controlled experiments in papers 2 and 3.

Subject B: Societal Challenges Posed by Digital Technologies

Topic 3: Explainable AI

This dissertation contributes to building responsible AI systems by addressing the problem of opaque ML models. It first investigates the preferences of XAI users and second develops and evaluates a process to guide the design of XAI systems to foster human agency by placing the user at the center of attention. Thus, for topic 3, this dissertation also employs the iterative cycle between deriving knowledge (BSR) and turning these insights into design and evaluation (DSR).

Both paper 4 and paper 5 build on an identical experimental setup that allows users to interact with XAI on a simplified task, employing the human-grounded evaluation scenario to assess the quality of the explanations from the users' perspective (Doshi-Velez and Kim 2018). The study is based on an ML-based smartphone application for plant species detection (cf. figure 4). Participants start by matching a leaf to a plant species. The following step presents the ML's matching recommendation and two different explanations justifying this recommendation. The participants choose their preferred explanation and justify it by selecting one or several reasons from a pre-defined list of explanation characteristics. Participants repeat this cycle multiple times with different samples and explanations.

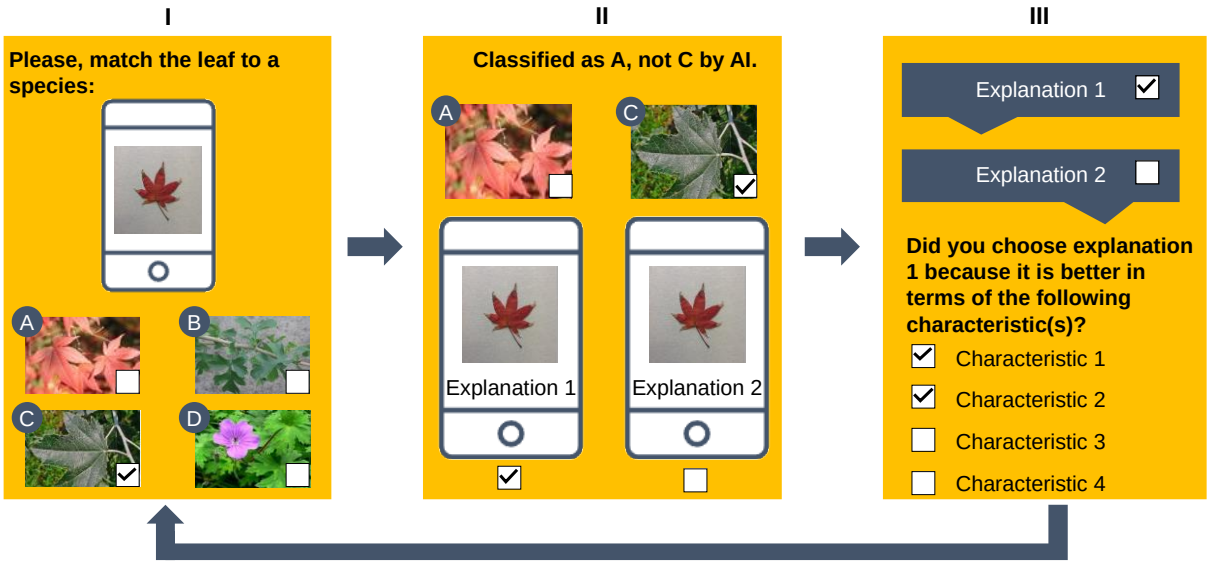


Figure 4: Experimental setup in papers 4 and 5.

Three guidelines for human-grounded evaluation inform the design of this experimental setup. First, it ensures rigor by using a real-world data set and a functionally complex AI system (Abdul et al. 2018). A neural network trained on a dataset of shape and texture attributes extracted from 340 images of leaf specimens from 30 different plant species provides the foundation for the application's recommendations (Silva 2013; Silva et al. 2013, 2014). Second, the AI employed in the experimental setup conducts image classification, one of the main tasks for today's AI applications (Whittaker et al. 2018). Thus, it presents a simplified task that is easily transferrable to real-world tasks (Doshi-Velez and Kim 2018). Finally, the scenario is accessible for laypeople while still not delivering obvious outputs that would render the explanations superfluous (Doshi-Velez and Kim 2018).

To evaluate explanations generated by XAI methods from the users' perspective and identify characteristics of explanations users appreciate in the context of XAI, paper 4 follows the BSR paradigm. The findings build on first a pre-study with 38 students aiming to complete the list of explanation characteristics and improve the study's comprehensibility. Second, the main study included 164 participants recruited via the platform Clickworker, chosen due to reduced threats to internal validity given similar data quality and results when compared with traditional methods (Buhrmester et al. 2011; Gürtzgen et al. 2018; Paolacci et al. 2010). Statistical tests determine if the share of a characteristic being decisive in all pairs of explanations was significantly greater. In addition, given the importance of "length," (Lombrozo 2007) the co-occurrence of characteristics with perceived length was analyzed.

Paper 5 follows the DSR paradigm to design a process that systematically guides the instantiation, calibration, and quality control of XAI systems such that they foster human agency and enable appropriate trust in AI systems. Both quantitative evidence that the artifact fulfills its objective and the theoretical foundations of its core concepts serve to demonstrate and evaluate the applicability and efficacy of the artifact in a realistic setting (Hevner et al. 2004). The findings are based on the experimental study described above (cf. figure 4). The main study was conducted in two iterations with 144 and 140 participants recruited via the platform Clickworker.

In turn, paper 6 employs the method of a systematic and structured literature review (Bandara et al. 2011; Webster and Watson 2002) to offer a comprehensive analysis and critical examination of XAI research in IS (Rowe 2014). Hevner highlights meta-artifacts such as a literature review synthesizing existing artifacts and processes as part of the *Rigor Cycle* in DSR, as it is vital for researchers to embed their artifacts in past knowledge to ensure their novelty (2007). The literature review follows three key steps: First, all relevant research sources are investigated (Webster and Watson 2002). Second, a search strategy is developed, including considerations regarding the appropriate time frame, applicable search terms, and fields (Cooper 1988; Levy and Ellis 2006). Third, all 168 articles identified as relevant are coded based on research concepts (Beese et al. 2019).

To provide an overview, table 2 summarizes the research paradigms, approaches and data sources discussed above.

Table 2: Overview of this dissertation's research paradigms, research approaches, and data.

Publication	Research Paradigm	Research Approach	Data
#JOBLESS #OLDER #DIGITAL – Digital Media User Types of the Older Unemployed	BSR	Investigation into the digital media preferences of older, unemployed individuals, building on a survey.	• Survey data
Activating Older Unemployed Individuals: A Case Study of Online Job Search Peer Groups	BSR	Investigation into the effectiveness of digital interventions for older, unemployed individuals, building on a controlled randomized field experiment.	• Survey data • Usage data
Leveraging the Power of Peer Groups for Refugee Integration: A Randomized Field Experiment Comparing Online and Offline Peer Groups	DSR	An online artifact to support refugee integration is developed and evaluated, including a comparative evaluation between the online and in-person realization, building on a controlled randomized field experiment.	• Survey data • Usage data • Demographic data (owned by public authority)
Evaluating Explainable Artificial Intelligence– What Users Really Appreciate	BSR	Investigation into characteristics of explanations XAI users appreciate based on a mixed-methods approach, including a human-grounded evaluation (user study).	• Publicly available data set for leaves and plant species • User study data

Fostering Human Agency: A Process for the Design of User-Centric XAI Systems	DSR	A novel process to guide the design of XAI toward human agency is developed and evaluated based on human-grounded evaluation (user study).	• Publicly available data set for leaves and plant species • User study data
Explainable Artificial Intelligence in Information Systems: A Review of the Status Quo and Future Research Directions	Literature Review	A comprehensive overview of XAI research in IS based on a systematic and structured literature review, including 168 research papers.	

1.4 Structure of the Dissertation

The dissertation is structured as follows: After the introduction, presenting this dissertation's motivation, and providing an overview of research objectives and methods, the individual papers of this dissertation are presented. Chapter 2 includes the papers addressing research on digital technologies to alleviate societal challenges (Subject A; Topics I and II). Chapter 3 comprises the papers addressing societal challenges posed by digital technologies (Subject B; Topic III).

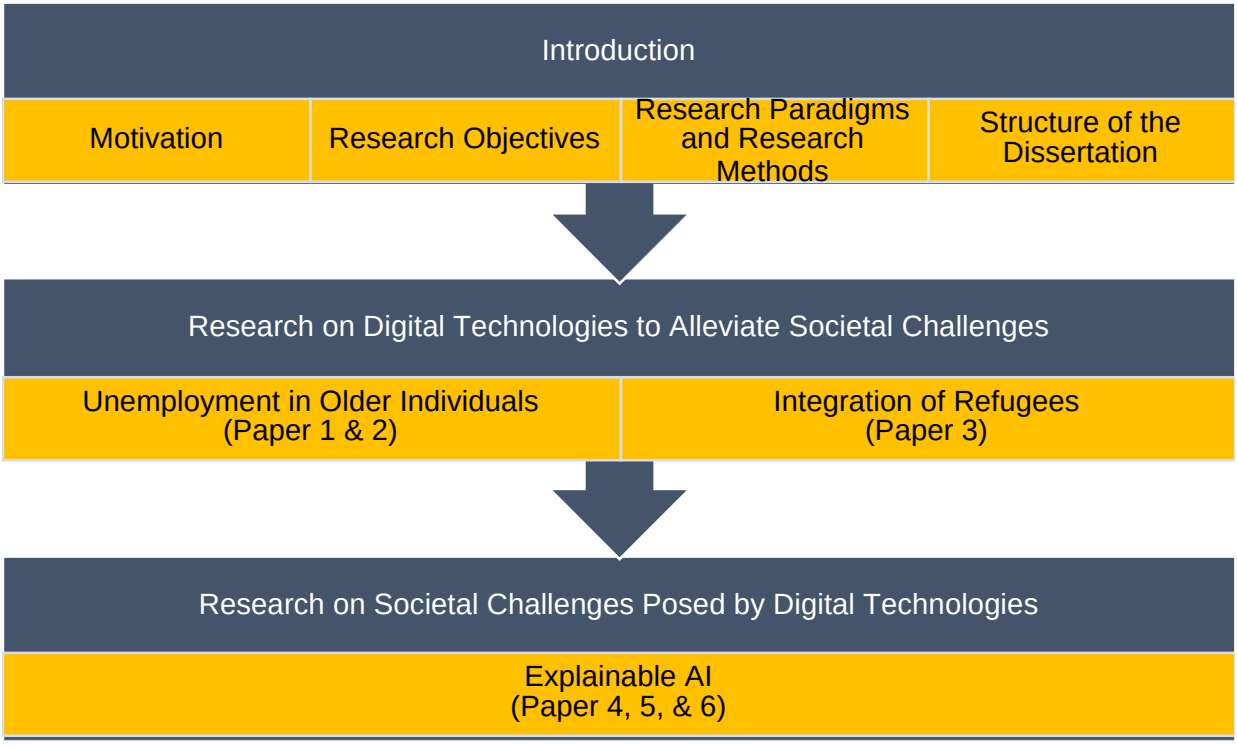


Figure 5: Overview of the structure of the dissertation.

1.5 References

Aanestad, M., Kankanhalli, A., Maruping, L., Pang, M. S., & Ram, S. (2022). Call for Papers MISQ Special Issue on Digital Technologies and Societal Justice. *MIS Quarterly*. https://misq.umn.edu/skin/frontend/default/misq/pdf/CurrentCalls/SI_DigitalTechnologies.pdf

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal, Canada, 1–18. https://doi.org/10.1145/3173574.3174156

Abedin, B., Junglas, I., Meske, C., Motahari-Nezhad, H. R., & Rabhi, F. (2020). Special Issue on: Designing and Managing Human-AI Interactions. *Information Systems Frontiers*. https://resource-cms.springernature.com/springer-cms/rest/v1/content/18030966/data/v3

AbuJarour, S., Köster, A., Krasnova, H., & Wiesche, M. (2021). Technology as a Source of Power: Exploring How ICT Use Contributes to the Social Inclusion of Refugees in Germany. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, Virtual Event, 2637–2646. https://doi.org/10.24251/HICSS.2021.322

AbuJarour, S., Krasnova, H., & Hoffmeier, F. (2018). ICT as an enabler: understanding the role of online communication in the social inclusion of Syrian refugees in Germany. In *Proceedings of the 26th European Conference on Information Systems*, Portsmouth, UK, 1-17.

AbuJarour, S., Wiesche, M., Díaz Andrade, A., Fedorowicz, J., Krasnova, H., Olbrich, S., Tan C-W., Urquhart C., & Venkatesh, V. (2019). ICT-enabled Refugee Integration: A Research Agenda. *Communications of the Association for Information Systems, 44*(1), 874–891. https://doi.org/10.17705/1CAIS.04440

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

Ager, A., & Strang, A. (2008). Understanding integration: A conceptual framework. *Journal of Refugee Studies, 21*(2), 166–191. https://doi.org/10.1093/jrs/fen016

Akar, E., Mardikyan, S., & Dalgic, T. (2019). User Roles in Online Communities and Their Moderating Effect on Online Community Usage Intention: An Integrated Approach. *International Journal of Human–Computer Interaction, 35*(6), 495–509. https://doi.org/10.1080/10447318.2018.1465325

Alencar, A. (2018). Refugee integration and social media: a local and experiential perspective. *Information, Communication & Society, 21*(11), 1588–1603. https://doi.org/10.1080/1369118X.2017.1340500

Alencar, A., Kondova, K., & Ribbens, W. (2019). The smartphone as a lifeline: an exploration of refugees' use of mobile communication technologies during their flight. *Media, Culture & Society, 41*(6), 828–844. https://doi.org/10.1177/0163443718813486

Allington, D., Duffy, B., Wessely, S., Dhavan, N., & Rubin, J. (2021). Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. *Psychological medicine, 51*(10), 1763-1769. https://doi.org/10.1017/S003329172000224X

Andresen, M. A. (2009). Asynchronous discussion forums: success factors, outcomes, assessments, and limitations. *Journal of Educational Technology & Society*, *12*(1), 249–257. http://www.jstor.org/stable/jeductechsoci.12.1.249

Ardakani, A. A., Kanafi, A. R., Acharya, U. R., Khadem, N., & Mohammadi, A. (2020). Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks. *Computers in biology and medicine, 121*, 103795. https://doi.org/10.1016/j.compbiomed.2020.103795

Arrieta, B. 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. https://doi.org/10.1016/j.inffus.2019.12.012

Bacishoga, K. B., & Johnston, K. A. (2013). Impact of mobile phones on integration: The case of refugees in South Africa. *The Journal of Community Informatics*, *9*(4), 1-20. https://doi.org/10.15353/joci.v9i4.3142

Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., McConnell, M. V., Percha, B., Snyder, T. M., & Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, *2*(1), 1-10. https://doi.org/10.1038/s41746-019-0105-1

Bandara, W., Miskon, S., & Fielt, E. (2011). A systematic, tool-supported method for conducting literature reviews in information systems. In *Proceedings of the 19th European Conference for Information Systems*, Helsinki, Finland, 1–13.

Bandura, A. (1997). *Self-efficacy: The exercise of control* (1st ed.). New York: Freeman.

Barak, A., Boniel-Nissim, M., & Suler, J. (2008). Fostering empowerment in online support groups. *Computers in Human Behavior, 24*(5), 1867–1883. doi:10.1016/j.chb.2008.02.004

Barnes, S. J., Bauer, H. H., Neumann, M. M., & Huber, F. (2007). Segmenting cyberspace: a customer typology for the internet. *European Journal of Marketing, 41*(1/2), 71–93. https://doi.org/10.1108/03090560710718120

Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Expl(AI)n It to Me – Explainable AI and Information Systems Research. *Business & Information Systems Engineering*, *63*(2), 79–82. https://doi.org/10.1007/s12599-021-00683-2

Beck, B. R., Shin, B., Choi, Y., Park, S., & Kang, K. (2020). Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Computational and structural biotechnology journal*, *18*, 784-790. https://doi.org/10.1016/j.csbj.2020.03.025

Beese, J., Haki, M. K., Aier, S., & Winter, R. (2019). Simulation-Based Research in Information Systems. Epistemic Implications and a Review of the Status Quo. *Business & Information Systems Engineering*, *61*(4), 503-521. https://doi.org/10.1007/s12599-018-0529-1

Belot, M., Kircher, P., & Muller, P. (2019). Providing Advice to Jobseekers at Low Cost: An Experimental Study on Online Advice. *Review of Economic Studies*, *86*(4), 1411–1447. https://doi.org/10.1093/restud/rdy059

Benbya, H., Nan, N., Tanriverdi, H., & Yoo, Y. (2020). Complexity and information systems research in the emerging digital world. *MIS Quarterly*, *44*(1), 1-17. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3539079

Bender, J. L., Katz, J., Ferris, L. E., & Jadad, A. R. (2013). What is the role of online support from the perspective of facilitators of face-to-face support groups? A multi-method study of the use of breast cancer online communities. *Patient Educ Couns, 93*(3), 472–479. https://doi.org/10.1016/j.pec.2013.07.009

Benton, M., & Glennie, A. (2016, October). *Digital humanitarianism: How Tech Entrepreneurs Are Supporting Refugee Integration*. Washington, DC: Migration Policy Institute. https://www.migrationpolicy.org/sites/default/files/publications/TCM-Asylum-Benton-FINAL.pdf

Betts, A., Sterck, O., Geervliet, R., MacPherson, C., Ali, A., & Memişoğlu, F. (2017). *Talent displaced: the economic lives of Syrian refugees in Europe.* Deloitte and the Refugee Studies Centre at the University of Oxford, Oxford. https://www2.deloitte.com/content/dam/Deloitte/global/Documents/About-Deloitte/talent-displaced-syrian-refugees-europe.pdf

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh J., Puri R., Moura J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 648-657. https://doi.org/10.1145/3351095.3375624

Biewen, M., Fitzenberger, B., Osikominu, A., & Waller, M. (2007). Which Program for Whom? Evidence on the Comparative Effectiveness of Public Sponsored Training Programs in Germany. *ZEW Discussion Paper*, No. 07- 042, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim. https://docs.iza.org/dp2885.pdf

Boockmann, B., & Brändle, T. (2019). Coaching, Counseling, Case-Working: Do They Help the Older Unemployed Out of Benefit Receipt and Back Into the Labor Market?. *German Economic Review*, *20*(4), e436–e468. https://doi.org/10.1111/geer.12174

Borzyskowski, I., Mazumder, A., Mateen, B., & Wooldridge, M. (2021). *AI and data science in the age of COVID-19.* The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2021-06/data-science-and-ai-in-the-age-of-covid_full-report_2.pdf

Braithwaite, D. O., Waldron, V. R., & Finn, J. (1999). Communication of social support in computer-mediated groups for people with disabilities. *Health Commun, 11*(2), 123–151. https://doi.org/10.1207/s15327027hc1102_2

Brandtzæg, P. B. (2010). Towards a unified Media-User Typology (MUT): A meta-analysis and review of the research literature on media-user typologies. *Computers in Human Behavior, 26*(5), 940–956. https://doi.org/10.1016/j.chb.2010.02.008

Briscese, G., Zanella, G., & Quinn, V. (2020). Improving Job Search Skills: A Field Experiment on Online Employment Assistance. *IZA Discussion Paper,* No. 13170, Institute for the Study of Labor (IZA), Bonn. https://www.iza.org/publications/dp/13170/improving-job-search-skills-a-field-experiment-on-online-employment-assistance

Buhrmester, M., Kwang T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science, 6*(3), 3–5. https://doi.org/10.1177/1745691610393980

Bynum, T. W. (2006). Flourishing ethics. *Ethics and Information Technology*, *8*, 157–173. https://doi.org/10.1007/s10676-006-9107-1

Card, D., Kluve, J., & Weber, A. (2018). What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations. *Journal of the European Economic Association*, *16*(3), 894-931. https://doi.org/10.1093/jeea/jvx028

Carr, N. (2008, February 21). How many computers does the world need? Fewer than you think. *The Guardian*. https://www.theguardian.com/technology/2008/feb/21/computing.supercomputers

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R. M., Kelley, T. D., Braines, D., Sensoy, M., Willis, C. J., & Gurram, P. (2017). Interpretability of deep learning models: A survey of results. In *IEEE Smart World, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, San Francisco, USA, 1–6. https://doi.org/10.1109/UIC-ATC.2017.8397411

Cheraghi-Sohi, S., Panagioti, M., Daker-White, G., Giles, S., Riste, L., Kirk, S., Ong, B. N., Poppleton, A., Campbell, S., & Sanders, C. (2020). Patient safety in marginalised groups: a narrative scoping review. *International Journal for Equity in Health*, *19*(26). https://doi.org/10.1186/s12939-019-1103-2

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society, 1*(104). https://doi.org/10.1007/BF03177550

Cotten, S. R., Ford, G., Ford, S., & Hale, T. M. (2012). Internet use and depression among older adults. *Computers in Human Behavior, 28*(2), 496–499. https://doi.org/10.1016/j.chb.2011.10.021

Cutrona, C. E., & Suhr, J. A. (1992). Controllability of Stressful Events and Satisfaction with Spouse Support Behaviors. *Communication Research*, *19*(2), 154–174. https://doi.org/10.1177/009365092019002002

Coulson, N. S. (2013). How do online patient support communities affect the experience of inflammatory bowel disease? An online survey. *JRSM Short Reports, 4*(8), 1–8. https://doi.org/10.1177/2042533313478004

Coussement, K., & Benoit, D. F. (2021). Interpretable data science for decision making. *Decision Support Systems*, *150*. https://doi.org/10.1016/j.dss.2021.113664

Dahya, N., & Dryden-Peterson, S. (2017). Tracing pathways to higher education for refugees: the role of virtual support networks and mobile phones for women in refugee camps. *Comparative Education*, *53*(2), 284-301. https://doi.org/10.1080/03050068.2016.1259877

Dar, A. B., Lone, A. H., Zahoor, S., Khan, A. A., & Naaz, R. (2020). Applicability of mobile contact tracing in fighting pandemic (COVID-19): issues, challenges and solutions. *Computer Science Review*, *38*, 100307. https://doi.org/10.1016/j.cosrev.2020.100307

Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dekker, R., Engbersen, G., Klaver, J., & Vonk, H. (2018). Smart Refugees: How Syrian Asylum Migrants Use Social Media Information in Migration Decision-Making. *Social Media + Society*, *4*(1), 1–11. https://doi.org/10.1177/2056305118764439

Diaz, A. A., & Doolin, B. (2016). Information and communication technology and the social inclusion of refugees. *MIS Quarterly*, *40*(2), 405-416. https://www.jstor.org/stable/26628912

Doran, D., Schulz, S., & Besold, T. R. (2018). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*, Bari, Italy. https://doi.org/10.48550/arXiv.1710.00794

Doshi-Velez, F., & Kim, B. (2018). Considerations for Evaluation and Generalization in Interpretable Machine Learning. In H. E. Jair, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. Gerven, (Eds.), *Explainable and Interpretable Models in Computer Vision and Machine Learning* (1st ed., 3–17). Springer. https://doi.org/10.1007/978-3-319-98131-4

Duryea, E. L., Adhikari, E. H., Ambia, A., Spong, C., McIntire, D., & Nelson, D. B. (2021). Comparison between in-person and audio-only virtual prenatal visits and perinatal outcomes. *JAMA Network Open*, *4*(4), 1-9. https://doi.org/ 10.1001/jamanetworkopen.2021.5854

Eide, E. (2020). Mobile Flight: Refugees and the Importance of Cell Phones. *Nordic Journal of Migration Research*, *10*(2), 67–81. http://doi.org/10.33134/njmr.250

Ekins, S., Mottin, M., Ramos, P. R. P. S., Sousa, B. K. P., Neves, B. J., Foil, D. H., Zorn, K. M., Braga, R. C., Coffee, M., Southan, C., Puhl, A. C., & Andrade, C. H. (2020). Déjà vu: stimulating open drug discovery for SARS-CoV-2. *Drug Discovery Today*, *25*(5), 928-941. https://doi.org/10.1016/j.drudis.2020.03.019

European Commission. (2021, April 21). *Proposal for a Regulation of the European Parliament and of the Council laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Parliament and the Council of the European Union. (2016, May 4). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

Farinango, C. D., Benavides, J. S., & Lopez, D. M. (2015). OpenUP/MMU-ISO 9241-210. Process for the Human Centered Development of Software Solutions, *IEEE Latin America Transactions*, *13*(11), 3668–3675. http://doi.org/10.1109/TLA.2015.7387947

Felgenhauer, A., Förster, M., Kaufmann, K., Klier, J., & Klier, M. (2019). Online peer groups – a design-oriented approach to addressing the unemployment of people with complex barriers. In *Proceedings of the 27th European Conference on Information Systems*, Stockholm, Sweden, 1-19.

Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., Parker, M., Bonsall, D., & Fraser, C. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, *368*(6491). http://doi.org/10.1126/science.abb6936

Feuls, M., Fieseler, C., Meckel, M., & Suphan, A. (2016). Being unemployed in the age of social media. *New Media & Society, 18*(6), 944–965. https://doi.org/10.1177/1461444814552637

Fugate, M., Kinicki, A. J., & Ashforth, B. E. (2004). Employability: A psycho-social construct, its dimensions, and applications. *Journal of Vocational Behavior*, *65*(1), 14–38. https://doi.org/10.1016/j.jvb.2003.10.005

Gaffield, C. (2020, April 17). COVID-19 is the First Global Pandemic in the Digital Age. *Royal Society of Canada – COVID 19 Series*. https://rsc-src.ca/en/voices/covid-19-is-first-global-pandemic-in-digital-age

Garg, R., & Telang, R. (2012). Role of Online Social Networks in Job Search by Unemployed Individuals. In *Proceeding of the 33rd International Conference on Information Systems*, Orlando, Florida, 1-15.

Garg, R., & Telang, R. (2018). To be or not to be linked: Online social networks and job search by unemployed workforce. *Management Science*, *64*(8), 3926-3941. https://doi.org/10.1287/mnsc.2017.2784

Goes, P. B. (2013). Editor's Comments: Information Systems Research and Behavioral Economics. *MIS Quarterly, 37*(3), iii-viii.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (1st ed.). MIT Press.

Google. (2019, May 8). People + AI Guidebook. *People + AI Research Team*. https://pair.withgoogle.com/guidebook

Goswami, S., Köbler, F., Leimeister, J. M., & Krcmar, H. (2010). Using online social networking to enhance social connectedness and social support for the elderly. In *Proceedings of the 31st International Conference on Information Systems*, St. Louis, USA, 1-10.

Grantz, K. H., Meredith, H. R., Cummings, D. A., Metcalf, C. J. E., Grenfell, B. T., Giles, J. R., Metha, S., Solomon, S., Labriqie, A., Kishore, N., Buckee, C. O., & Wesolowski, A. (2020). The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications*, *11*(1), 1-8. https://doi.org/10.1038/s41467-020-18190-5

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly,* 37(2), 337-355. https://www.jstor.org/stable/43825912

Guidotti, R., Monreale, A., Matwin, S., & Pedreschi, D. (2020). Black Box Explanation by Learning Image Exemplars in the Latent Feature Space. In U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, & C. Robardet, (Eds.), *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Lecture Notes in Computer Science* (vol 11906, 189-205). Springer, Cham. https://doi.org/10.1007/978-3-030-46150-8_12

Gürtzgen, N., Nolte, A., Pohlan, L., & Berg, G. J. (2018). Do digital information technologies help unemployed job seekers find a job? Evidence from the Internet Expansion in Germany. *IZA Discussion Paper*, No. 11555, Institute for the Study of Labor (IZA), Bonn. https://www.iza.org/publications/dp/11555/do-digital-information-technologies-help-unemployed-job-seekers-find-a-job-evidence-from-the-broadband-internet-expansion-in-germany

Harder, N., Figueroa, L., Gillum, R. M., Hangartner, D., Laitin, D. D., & Hainmueller, J. (2018). Multidimensional measure of immigrant integration. *PNAS*, *115*(45),11483-11488. https://doi.org/10.1073/pnas.1808793115

Heaven, W. D. (2021, July 30). Hundreds of AI tools have been built to catch covid. None of them helped. *MIT Technology Review*. https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

Helsper, E. J. (2011, July). *The Emergence of a Digital Underclass: Digital Policies in the UK and Evidence for Inclusion.* LSE Media Policy Project Series, Media Policy Brief 3. London School of Economics and Political Science Department of Media and Communications. http://eprints.lse.ac.uk/38615/1/LSEMPPBrief3.pdf

Helsper, E. J., & Reisdorf, B. C. (2017). The emergence of a "digital underclass" in Great Britain and Sweden: Changing reasons for digital exclusion. *New Media and Society, 19*(8), 1253–1270. https://doi.org/10.1177/1461444816634676

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*(2), 1-6. http://aisel.aisnet.org/sjis/vol19/iss2/4

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly, 28*(1), 75–105. https://doi.org/10.2307/25148625

Hill, R., Betts, L., & Gardner, S. (2015). Empowerment and enablement through digital technology in the generation of the digital age. *Computers in Human Behavior*, *48*, 1–23. https://doi.org/10.1016/j.chb.2015.01.062

Hilton, D. J., & Erb, H. P. (1996). Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance. *Thinking & Reasoning, 2*(4), 273–308. https://doi.org/10.1080/135467896394447

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-Based Causal Attribution. The Abnormal Conditions Focus Model. *Psychological Review, 93*(1), 75–88. https://doi.org/10.1037/0033-295X.93.1.75

HLEG-AI. (2019, April 8). *Ethics Guidelines for Trustworthy Artificial Intelligence*. Brussels: Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

Hunsaker, A., & Hargittai, E. (2018). A review of Internet use among older adults. *New Media & Society, 20*(10), 3937–3954. https://doi.org/10.1177/1461444818787348

Hynie, M., Korn, A., & Tao, D. (2016). Social context and social integration for government assisted refugees in Ontario, Canada. In M. Poteet, & S. Nourpanah, (Eds.), *After the flight: the dynamics of refugee settlement and integration* (183-227). Cambridge Scholars Publishing, Cambridge.

IDEO. (2015, January 1). The Field Guide to Human-Centered Design. *IDEO.org*. https://www.designkit.org//resources/1

International Telecommunication Union. (2021, January). *Measuring Digital Development. Facts and Figures 2021*. ITU Publications. https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2021.pdf

ISO. (2019, July). *ISO 9241-210. 2019 - Ergonomics of Human-System Interaction — Part 210: Human-Centred Design for Interactive Systems*, Geneva: International Organization for Standardization. https://www.iso.org/standard/77520.html

Jebb, A. T., Morrison M., Tay, L., & Diener, E. (2020). Subjective Well-Being Around the World: Trends and Predictors Across the Life Span. *Psychological Science*, *31*(3), 293–305. https://doi.org/10.1177/0956797619898826

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021). How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations. In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* Virtual Event, 805-815. https://doi.org/10.1145/3442188.3445941

Jirotka, M., & Stahl, B. C. (2020). The need for responsible technology. *Journal of Responsible Technology*, *1*, 100002. https://doi.org/10.1016/j.jrt.2020.100002

Johns Hopkins University. (2022). *Coronavirus resource center: COVID-19 Dashboard.* Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. https://coronavirus.jhu.edu/map.html

Kahneman, D. & Tversky, A. (1981). The Simulation Heuristic. In D. Kahneman, P. Slovic, & A. Tversky, (Eds.), *Judgement under uncertainty: Heuristics and biases* (1st ed., 201-208). New York: Cambridge University.

Katz, A. H., & Bender, E. L. (1976). *The strength in us: self-help groups in the modern world* (1st ed.). New York: New Viewpoints.

Kaufmann, K. (2018). Navigating a new life: Syrian refugees and their smartphones in Vienna. *Information, Communication & Society, 21*(6), 882–898. https://doi.org/10.1080/1369118X.2018.1437205

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V. A. N., Wen, C., Zhang, E. D., Zhang, C. L., Li, O., Wang, X., Singer, M. A., Sun, X., Xu, J., Tafreshi, A., Lewis, M. A., Xia, H., & Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell, 172*(5), 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

Kiesler, S., Zubrow, D., Moses, A. M., & Geller, V. (1985). Affect in Computer-Mediated Communication: An Experiment in Synchronous Terminal-to-Terminal Discussion. *Human-Computer Interaction, 1*(1), 77–104. https://doi.org/10.1207/s15327051hci0101_3

Kim, J. Y. (2018). A study of social media users' perceptional typologies and relationships to self- and personality. *Internet Research, 28*(3), 767-784. https://doi.org/10.1108/IntR-05-2017-0194

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. In *30th Conference on Neural Information Processing Systems (NIPS 2016),* Barcelona, Spain, 1-9.

Kirsch, A. (2018). Explain to whom? Putting the user in the center of explainable AI. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*. Bari, Italy.

Klehe, U., Koen, J., & DePater, I. E. (2012). Ending on the scrap heap: The experience of job loss and job search among older workers. In J. W. Hedge, & W. C. Borman, (Eds.), *The Oxford handbook of work and aging* (313–340). Oxford University Press.

Klier, J., Klier, M., Thiel, L., & Agarwal, R. (2019). Power of Mobile Peer Groups: A Design-Oriented Approach to Address Youth Unemployment. *Journal of Management Information Systems*, *36*(1), 158–193. https://doi.org/10.1080/07421222.2018.1550557

Knowles, T. (2021, July 13). AI will have a bigger impact than fire, says Google boss Sundar Pichai. *The Times*. https://www.thetimes.co.uk/article/ai-will-have-a-bigger-impact-than-fire-says-google-boss-sundar-pichai-rk8bdst7r

König, R., Seifert, A., & Doh, M. (2018). Internet use among older Europeans: an analysis based on SHARE data. *Universal Access in the Information Society, 17*(3), 621–633. https://doi.org/10.1007/s10209-018-0609-5

Kouzy, R., Jaoude, J. A., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., & Baddour, K. (2020). Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus*, *12*(3). https://doi.org/ 10.7759/cureus.7255

Kuhn, P., & Mansour, H. (2014). Is Internet Job Search Still Ineffective?. *The Economic Journal*, *124*(581), 1213–1233. https://doi.org/10.1111/ecoj.12119

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning,* Virtual Event, 5491-5500.

Kutscher, N., & Kreß, L. M. (2016). "Internet is the same like food"–An empirical study on the use of digital media by unaccompanied minor refugees in Germany. *Transnational Social Review*, *6*(1-2), 200-203. https://doi.org/10.1080/21931674.2016.1184819

Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals, 139,* 110059. https://doi.org/10.1016/j.chaos.2020.110059

Lee, B., Chen, Y., & Hewitt, L. (2011). Age differences in constraints encountered by seniors in their use of computers and the internet. *Computers in Human Behavior*, *27*(3), 1231–1237. https://doi.org/10.1016/j.chb.2011.01.003

Legner, C., Eymann, T., Hess, T., Matt, C., Böhmann, T., Drews, P., Mädche, A., Urbach, N., & Ahlemann, F. (2017). Digitalization: opportunity and challenge for the business and information systems engineering community. *Business & Information Systems Engineering*, *59*(4), 301-308. https://doi.org/10.1007/s12599-017-0484-2

Levy, Y., & Ellis, T. J. (2006). A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. *Informing Science Journal, 9*, 181-212.

Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., Cao, K., Liu, D., Wang, G., Xu, Q., Fang, X., Zhang, S., Xia, J., & Xia, J. (2020). Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. *Radiology*, *296*(2), E65–E71. https://doi.org/10.1148/radiol.2020200905

Liao, S. M. (2020). *Ethics of Artificial Intelligence* (1st ed.). Oxford University Press.

Lipton, P. (2000). Inference to the Best Explanation. In W. H. Newton-Smith (Ed.), *A Companion to the Philosophy of Science* (1st ed., 184-193). Blackwell.

Liu, S., Huang, J. L., & Wang, M. (2014). Effectiveness of Job Search Interventions: A Meta-Analytic Review. *Psychological Bulletin*, *140*(4), 1–33. https://doi.org/10.1037/a0035923

Liu, J. L., & Wyatt, J. C. (2011). The case for randomized controlled trials to assess the impact of clinical information systems. *Journal of the American Medical Informatics Association*, *18*(2), 173-180. https://doi.org/10.1136/jamia.2010.010306

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257. https://doi.org/10.1016/j.cogpsych.2006.09.006

Lombrozo, T. (2012). Explanation and Abductive Inference. In K. J. Holyoak, & R. G. Morrisson, (Eds.), *The Oxford Handbook of Thinking and Reasoning* (1st ed., 260–276). Oxford University Press.

Lüders, M., & Brandtzæg, P. B. (2017). 'My children tell me it's so simple': A mixed-methods approach to understand older non-users' perceptions of Social Networking Sites. *New Media & Society*, *19*(2), 181–198. https://doi.org/10.1177/1461444814554064

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 1-10.

Majchrzak, A., Markus, M. L., & Wareham, J. (2014). Call for Papers MISQ Special Issue on ICT and Societal Challenges. *MIS Quarterly*. https://misq.umn.edu/skin/frontend/default/misq/pdf/CurrentCalls/ICTChallenges.pdf

Majchrzak, A., Markus, M. L., & Wareham, J. (2016). Designing for digital transformation: Lessons for information systems research from the study of ICT and societal challenges. *MIS Quarterly*, *40*(2), 267-277. https://www.jstor.org/stable/26628906

Marbán, O., Segovia, J., Menasalvas, E., & Fernández-Baizán, C. (2009). Toward Data Mining Engineering: A Software Engineering Approach. *Information Systems*, *34*(1), 87–107. https://doi.org/10.1016/j.is.2008.04.003

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez Orallo, J., Kull, M., Lachiche, N., Ramirez Quintana, M. J., & Flach, P. A. (2019). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048 – 3061. https://doi.org/ 10.1109/TKDE.2019.2962680

Maurer, T. J. (2001). Career-relevant learning and development, worker age, and beliefs about self-efficacy for development. *Journal of Management*, *27*(2), 123–140. https://doi.org/10.1177/014920630102700201

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12. https://doi.org/10.1609/aimag.v27i4.1904

McCloy, R., & Byrne, R. M. J. (2000). Counterfactual thinking about controllable events. *Memory & Cognition*, *28*(6), 1071–1078. https://doi.org/10.3758/BF03209355

McClure, J. (2002). Goal-based Explanations of Actions and Outcomes. *European Review of Social Psychology*, *12*(1), 201–235. https://doi.org/10.1080/14792772143000067

McGregor, S. (2020, November 18). When AI Systems Fail: Introducing the AI Incident Database. *Partnership on AI*. https://partnershiponai.org/aiincidentdatabase/

McQuaid, R. W. (2006). Job search success and employability in local labor markets. *The Annals of Regional Science*, *40*(2), 407–421. https://doi.org/10.1007/s00168-006-0065-7

McQuaid, R. W., Lindsay, C., & Greig, M. (2004). 'Reconnecting' the Unemployed Information and communication technology and services for jobseekers in rural areas. *Information, Communication & Society*, *7*(3), 364–388. https://doi.org/10.1080/1369118042000284605

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35. https://doi.org/10.1145/3457607

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2020). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, *39*(1), 53-63. https://doi.org/10.1080/10580530.2020.1849465

Meske, C., Abedin, B., Junglas, I., & Rabhi, F. (2021). Introduction to the Minitrack on Explainable Artificial Intelligence (XAI). In *Proceedings of the 54th Hawaii International Conference on System Sciences*, Virtual Event. https://aisel.aisnet.org/hicss-54/da/xai/1/

Meske, C., Abedin, B., Klier, M., & Rabhi, F. (2022). CfP special issue on "Explainable and responsible artificial intelligence". *Electronic Markets*. http://www.electronicmarkets.org/call-for-papers/single-view-for-cfp/datum/2021/04/29/cfp-special-issue-on-explainable-and-responsible-artificial-intelligence/

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*,1–38. https://doi.org/10.1016/j.artint.2018.07.007

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-17 Workshop on Explainable AI (XAI) Proceedings*, Melbourne, Australia, 36–42. https://doi.org/10.48550/arXiv.1712.00547

Miller, D. T., & Gunasegaram, S. (1990). Temporal Order and the Perceived Mutability of Events: Implications for Blame Assignment. *Journal of Personality and Social Psychology*, *59*(6), 1111–1118. https://doi.org/10.1037/0022-3514.59.6.1111

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

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal, 73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Morris, A., Goodman, J., & Brading, H. (2007). Internet use and non-use: Views of older users. *Universal Access in the Information Society*, *6*(1), 43–57. https://doi.org/10.1007/s10209-006-0057-5

Ngan, H. Y., Lifanova, A., Jarke, J., & Broer, J. (2016). Refugees Welcome: Supporting Informal Language Learning and Integration with a Gamified Mobile Application. In K. Verbert, M. Sharples, & T. Klobučar, (Eds.), *Adaptive and Adaptable Learning. EC-TEL 2016. Lecture Notes in Computer Science* (vol 9891, 521-524). Springer, Cham. https://doi.org/10.1007/978-3-319-45153-4_54

Niehaves, B., & Becker, J. (2008). The Age-Divide in E-Government — Data, Interpretations, Theory Fragments. In M. Oya, R. Uda, & C. Yasunobu, (Eds.), *Towards Sustainable Society on Ubiquitous Networks. IFIP – The International Federation for Information Processing*, (vol 286, 279 - 287). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-85691-9_24

Niela-Vilén, H., Axelin, A., Salanterä, S., & Melender, H. L. (2014). Internet-based peer support for parents: A systematic integrative review. *International Journal of Nursing Studies, 51*(11), 1524- 1537. https://doi.org/10.1016/j.ijnurstu.2014.06.009

Norman, D. A., & Draper, S. W. (1986). *User Centered System Design: New Perspectives on Human-Computer Interaction* (1st ed.). CRC Press.

OECD. (2019, August 30). *Working Better with Age*. Ageing and Employment Policies, OECD Publishing, Paris. https://doi.org/10.1787/c4d4f66a-en

Olphert, C., Damodaran, L., & May, A. (2005). Towards digital inclusion - engaging older people in the 'digital world.' In *Accessible Design in the Digital World Conference 2005 (AD)*, Dundee, Scotland, 1-7. https://doi.org/10.14236/ewic/AD2005.17

Ordonez, T. N., Yassuda, M. S., & Cachioni, M. (2011). Elderly online: Effects of a digital inclusion program in cognitive performance. *Archives of Gerontology and Geriatrics*, *53*(2), 216–219. https://doi.org/10.1016/j.archger.2010.11.007

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411–419. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1626226

Prevatt, B. S., Lowder, E. M., & Desmarais, S. L. (2018). Peer-support intervention for postpartum depression: participant satisfaction and program effectiveness. *Midwifery*, *64*, 38–47. https://doi.org/10.1016/j.midw.2018.05.009

Rader, E., & Gray, R. (2015). Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, Seoul, Republic of Korea, 173-182. https://doi.org/10.1145/2702123.2702174

Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? A proposal of user-centered explainable AI. In *Joint Proceedings of the ACM IUI 2019 Workshops,* Los Angeles, USA.

Rife, J. C., & Belcher, J. R. (1994). Assisting Unemployed Older Workers to Become Reemployed: An Experimental Evaluation. *Research on Social Work Practice*, *4*(1), 3–13. https://doi.org/10.1177/104973159400400101

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, R. J., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Rudd, J. H. F., Sala, E., & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, *3*(3), 199-217. https://doi.org/10.1038/s42256-021-00307-0

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science, 7*(10), 201199. https://doi.org/10.1098/rsos.201199

Rowe, F. (2014). What literature review is not: diversity, boundaries and recommendations. *European Journal of Information Systems, 23*(3), 241–255. https://doi.org/10.1057/ejis.2014.7

Rupert, D. J., Poehlman, J. A., Hayes, J. J., Ray, S. E., & Moultrie, R. R. (2017). Virtual versus in-person focus groups: Comparison of costs, recruitment, and participant logistics. *Journal of Medical Internet Research*, *19*(3), e80. https://doi.org/10.2196/jmir.6980

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Rzepka, C., & Berger, B. (2018). User Interaction with AI-enabled Systems: A systematic review of IS research. In *Proceedings of the 39th International Conference on Information Systems,* San Francisco, USA, 1-17.

Sachs, J. D., Modi, V., Figueroa, H., Machado, M., Sanyal, K., Khatun, F., & Reid, K. (2016). *How information and communications technology can accelerate action on the sustainable development goals*. The Earth Institute at Columbia University. https://www.ericsson.com/assets/local/news/2016/05/ict-sdg.pdf

Sannomiya, M., & Kawaguchi, A. (1999). Cognitive Characteristics Mediated Communication in Group Discussion: An Examination from Three Dimensions. *Educational Technology Research*, *22*(1-2), 19–25. https://doi.org/10.15077/etr.KJ00003899161

Schäfer-Siebert, K., & Verhalen, N. (2021). Thanks for your help!–the value of Q&A websites for refugee integration. In *Proceedings of the 54th Hawaii International Conference on System Sciences,* Virtual Event, 2647-2656. https://doi.org/10.24251/HICSS.2021.323

Schreieck, M., Wiesche, M., & Krcmar, H. (2017a). Governing nonprofit platform ecosystems–an information platform for refugees. *Information Technology for Development*, *23*(3), 618-643. https://doi.org/10.1080/02681102.2017.1335280

Schreieck, M., Zitzelsberger, J., Siepe, S., Wiesche, M., & Krcmar, H. (2017b). Supporting Refugees in Everyday Life-Intercultural Design Evaluation of an Application for Local Information. In *Twenty First Pacific Asia Conference on Information Systems*, Langkawi, Malaysia, 1-12.

Schmidt, C. (2007). Wirkungsorientierte Evaluation in der beruflichen Rehabilitation, *IQPR Forschungsbericht Nr. 5*, Cologne.

Shapley, L. (1953). A Value for n-Person Games. In H. Kuhn, & A. Tucker, (Eds.), *Contributions to the Theory of Games II* (307-317). Princeton University Press. https://doi.org/10.1515/9781400881970-018

Sheff, D. (1985, February 1). Playboy Interview: Steve Jobs. *Playboy*. https://www.scribd.com/doc/43945579/Playboy-Interview-With-Steve-Jobs

Siddiquee, A., & Kagan, C. (2006). The internet, empowerment, and identity: an exploration of participation by refugee women in a Community Internet Project (CIP) in the United Kingdom (UK). *Journal of Community & Applied Social Psychology*, *16*(3), 189-206. https://doi.org/10.1002/casp.855

Silva, P. F. B. (2013). *Development of a System for Automatic Plant Species Recognition*. https://hdl.handle.net/10216/67734

Silva, P. F. B., Marçal, A. R. S., & da Silva, R. M. A. (2013). Evaluation of Features for Leaf Discrimination. In M. Kamel, & A. Campilho, (Eds.), *Image Analysis and Recognition. ICIAR 2013. Lecture Notes in Computer Science*, (vol 7950, 197–204). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39094-4_23

Silva, P. F. B., Marçal, A. R. S., & da Silva, R. M. A. (2014). Leaf Dataset. *UCI Machine Learning Repository*. https://archive.ics.uci.edu/ml/datasets/Leaf

Siren, A., & Knudsen, S. G. (2017). Older adults and emerging digital service delivery: A mixed methods study on information and communications technology use, skills, and attitudes. *Journal of aging & social policy*, *29*(1), 35-50. https://doi.org/10.1080/08959420.2016.1187036

Stahl, B. C., Eden, G., Jirotka, M., & Coeckelbergh, M. (2014). From computer ethics to responsible research and innovation in ICT: The transition of reference discourses informing ethics-related research in information systems. *Information & Management*, *51*(6), 810-818. https://doi.org/10.1016/j.im.2014.01.001

Still, B., & Crane, K. (2017). *Fundamentals of User-Centered Design: A Practical Approach, Boca Raton* (1st ed.). FL: CRC Press. https://doi.org/10.4324/9781315200927

Suphan, A., Feuls, M., & Fieseler, C. (2012). Social Media's Potential in Improving the Mental Well-Being of the Unemployed. In K. Eriksson-Backa, A. Luoma, & E. Krook, (Eds.), *Exploring the Abyss of Inequalities. WIS 2012. Communications in Computer and Information Science* (vol 313, 10-28). https://doi.org/10.1007/978-3-642-32850-3_2

Sun, L., Song, F., Shi, N., Liu, F., Li, S., Li, P., Zhang, W., Jiang, X., Zhang, Y., Sun, L., Chen, X., & Shi, Y. (2020). Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19. *Journal of Clinical Virology*, *128,* 104431. https://doi.org/10.1016/j.jcv.2020.104431

Suri, T., & Jack, W. (2016). The long run poverty and gender impacts of mobile money. *Science*, *354*(6317), 1288–1292. https://doi.org/10.1126/science.aah5309

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435-467. https://doi.org/10.1017/S0140525X00057046

Tisch, A. (2015). The employability of older job-seekers: Evidence from Germany. *Journal of the Economics of Ageing*, *6*, 102–112. https://doi.org/10.1016/j.jeoa.2014.07.001

United Nations. (2019, June 10). *The Age of Digital Interdependence*. Report of the UN Secretary-General's High-level Panel on Digital Cooperation. https://www.un.org/en/pdfs/DigitalCooperation-report-for%20web.pdf

United Nations. (2020). *World Population Ageing 2020 Highlights: Living arrangements of older persons*. Department of Economic and Social Affairs, Population Division. https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa_pd-2020_world_population_ageing_highlights.pdf

United Nations. (2021, November 25). *193 countries adopt first-ever global agreement on the Ethics of Artificial Intelligence*. UN News. https://news.un.org/en/story/2021/11/1106612

United Nations General Assembly. (1951, July 28). *Convention Relating to the Status of Refugees*. United Nations Treaty Series. https://treaties.un.org/pages/ViewDetailsII.aspx?src=TREATY&mtdsg_no=V-2&chapter=5&Temp=mtdsg2&clang=_en

United Nations General Assembly. (2015, September 25). *Transforming our World: The 2030 Agenda for Sustainable Development*. https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

United Nations High Commissioner for Refugees. (2013, September). *A new beginning: refugee integration in Europe*. United Nations High Commissioner for Refugees. https://www.unhcr.org/protection/operations/52403d389/new-beginning-refugee-integration-europe.html

United Nations High Commissioner for Refugees. (2016, September). *Connecting refugees: how internet and mobile connectivity can improve refugee well-being and transform humanitarian action.* https://www.unhcr.org/5770d43c4.pdf

United Nations High Commissioner for Refugees. (2020, December). *Global trends: forced displacement in 2019*. United Nations High Commissioner for Refugees. https://www.unhcr.org/flagship-reports/globaltrends/

United Nations High Commissioner for Refugees. (2021, November). *Refugee Data Finder*. Refugee Population Statistics Database. https://www.unhcr.org/refugee-statistics/

United Nations High Commissioner for Refugees. (2022, May 13). *Ukraine Refugee Situation*. Operational Data Portal. https://data2.unhcr.org/en/situations/ukraine

Urbach, N., Smolnik, S., & Riempp, G. (2009). The State of Research on Information Systems Success. *Business & Information Systems Engineering, 1*(4), 315-325. https://doi.org/10.1007/s12599-009-0059-y

van der Waa, J., Robeer, M., van Diggelen, J., Brinkhuis, M., & Neerincx, M. (2018). Contrastive Explanations with Local Foil Trees. In *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning*, Stockholm, Sweden, 41–46. https://doi.org/10.48550/arXiv.1806.07470

Vansteenkiste, S., Deschacht, N., & Sels, L. (2015). Why are unemployed aged fifty and over less likely to find a job? A decomposition analysis. *Journal of Vocational Behavior*, 90, 55–65. https://doi.org/10.1016/j.jvb.2015.07.004

Venkatesh, V., Rai, A., Sykes, T. A., & Aljafari, R. (2016). Combating Infant Mortality in Rural India: Evidence from a Field Study of eHealth Kiosk Implementations. *MIS Quarterly*, *40*(2), 353-380. https://www.jstor.org/stable/26628910

Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 910-932. https://doi.org/10.1109/JSTSP.2020.3002101

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841– 887.

Wanberg, C. R., Hough, L. M., & Song, Z. (2002). Predictive Validity of a Multidisciplinary Model of Reemployment Success. *Journal of Applied Psychology*, *87*(6), 1100–1120. https://doi.org/10.1037/0021-9010.87.6.1100

Wang, X., Parameswaran, S., Bagul, D. M., & Kishore, R. (2017). Does online social support work in stigmatized chronic diseases? A study of the impacts of different facets of informational and emotional support on self-care behavior in an HIV online forum. In *Proceedings of the 38th International Conference on Information Systems*, Seoul, South Korea, 1-19.

Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. Christchurch, New Zealand, 109–116. https://doi.org/10.1109/HRI.2016.7451741

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. In *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, 1–15. https://doi.org/10.1145/3290605.3300831

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii-xxiii. https://www.jstor.org/stable/4132319

Weisberg, D. S., Taylor, J. C. V. & Hopkins, E. J. (2015). Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision Making*, *10*(5), 429–441.

Whitelaw, S., Mamas, M. A., Topol, E., & Van Spall, H. G. (2020). Applications of digital technology in COVID-19 pandemic planning and response. *The Lancet Digital Health*, *2*(8), e435-e440. https://doi.org/10.1016/S2589-7500(20)30142-4

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2018, December). *AI Now Report 2018*. AI Now Institute. https://ainowinstitute.org/AI_Now_2018_Report.pdf

Wilde, T., & Hess, T. (2007). Forschungsmethoden der Wirtschaftsinformatik. *Wirtschaftsinformatik*, *49*(4), 280-287. https://doi.org/10.1007/s11576-007-0064-z

World Economic Forum. (2021, January 19). *The Global Risk Report 2021*. World Economic Forum Global Risk Initiative. https://www.weforum.org/reports/the-global-risks-report-2021

World Health Organization. (2020, March 12). *WHO announces COVID-19 outbreak as a pandemic*. Health Topics. https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Dahly, D. D. L., Damen, J. A., Debray, T. P. A., de Jong, V. M. T., Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Heus, P., Kammer, M., Kreuzberger, N., … & van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*, *369*. https://doi.org/10.1136/bmj.m1328

Xue, A. (2021). End-to-end chinese landscape painting creation using generative adversarial networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Virtual Event, 3863-3871.

Yan, L., Zhang, H., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L. Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., ... & Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence*, 2, 283–288. https://doi.org/10.1038/s42256-020-0180-7

Yin, R. K. (2013). Validity and generalization in future case study evaluations. *Evaluation*, *19*(3), 321-332. https://doi.org/10.1177/1356389013497081

Zawacki-Richter, O., Müskens, W., Krause, U., Alturki, U., & Aldraiweesh, A. (2015). Student Media Usage Patterns and Non-Traditional Learning in Higher Education. *International Review of Research in Open and Distributed Learning*, *16*(2), 136–170. https://doi.org/10.19173/irrodl.v16i2.1979

Zeroual, A., Harrou, F., Dairi, A., & Sun, Y. (2020). Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos, Solitons & Fractals*, *140*, 110121. https://doi.org/10.1016/j.chaos.2020.110121

Zhang, W., Xiang, L., Jia, X.-D., Ma, H., Luo, Z., & Li, X. (2020). Machinery fault diagnosis with imbalanced data using deep generative adversarial networks. *Measurement*, *152*, 107377. https://doi.org/10.1016/j.measurement.2019.107377

Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digital Medicine*, *4*(1), 1-5. https://doi.org/10.1038/s41746-020-00372-6

2. Research on Digital Technologies to Alleviate Societal Challenges

2.1 #JOBLESS #OLDER #DIGITAL – DIGITAL MEDIA USER TYPES OF THE OLDER UNEMPLOYED

Full Citation:	Klier, Julia; Klier, Mathias; Schäfer-Siebert, Katharina; & Sigler, Irina (2020). #JOBLESS #OLDER #DIGITAL – DIGITAL MEDIA USER TYPES OF THE OLDER UNEMPLOYED. In *Proceedings of the 28th European Conference on Information Systems*, Virtual Event, 1-17. https://aisel.aisnet.org/ecis2020_rp/206
Copyright Note:	Reprinted according to author's rights.

6-15-2020

#JOBLESS #OLDER #DIGITAL – DIGITAL MEDIA USER TYPES OF THE OLDER UNEMPLOYED

Julia Klier
University of Regensburg, julia.klier@wiwi.uni-regensburg.de

Mathias Klier
University of Ulm, mathias.klier@uni-ulm.de

Katharina Schäfer-Siebert
University of Ulm, katharina.schaefer-siebert@uni-ulm.de

Irina Sigler
University of Ulm, irina.hardt@uni-ulm.de

#JOBLESS #OLDER #DIGITAL – DIGITAL MEDIA USER TYPES OF THE OLDER UNEMPLOYED

Research paper

Klier, Julia, University of Regensburg, Regensburg, Germany, julia.klier@ur.de

Klier, Mathias, University of Ulm, Ulm, Germany, mathias.klier@uni-ulm.de

Schäfer-Siebert, Katharina, University of Ulm, Ulm, Germany, katharina.schaefer-siebert@uni-ulm.de

Sigler, Irina, University of Ulm, Ulm, Germany, irina.sigler@uni-ulm.de

Abstract

The double burden of age and unemployment suggests that older unemployed individuals are particularly affected by digital divide. To investigate the characteristics and user types of older unemployed individuals with respect to digital media, we conduct a survey on this target group in cooperation with the German Federal Employment Agency. We apply cluster analysis to give more nuanced insights into people's usage of digital media services for job search, their attitude with respect to those services, currently perceived barriers and product preferences. Results suggest that older unemployed individuals are a nuanced group with a great range of different usage behaviours and attitudes towards digital media. Specifically, we identify the user types Digital Sceptics, Digital Classics, Digital Interactives and Digital Contributors. We show that a large group of older unemployed individuals is accessible for digital media in the context of job search and identify pathways of how our findings can be used to leverage digital media use for job search in this target group. Our study contributes to literature on digital media user typologies and has important policy implications for addressing the digital divide.

Keywords: User Typologies, Unemployment, Digital Divide, Cluster Analysis.

1 Introduction

Governments across the globe are increasingly using Information and Communication Technology (ICT) as a catalyst for more effective, efficient, and democratic public services (United Nations, 2018). However, to date, there is still a lack of adoption among citizens (Helbig et al., 2009; United Nations, 2018). In particular, marginalized groups show lower usage rates of digital public services (Niehaves and Becker, 2008; United Nations, 2018). To reduce the risk of further marginalization through the digitalisation of public services, regulators are required to take action (Alshibly and Chiong, 2015; Bertot et al., 2016; United Nations, 2018).

Two groups in society seem to be particularly affected by this risk: older people and unemployed individuals. Older people do not only show a low uptake of digital public services (Niehaves and Becker, 2008; United Nations, 2018) but also lack behind in overall digital media usage (Hunsaker and Hargittai, 2018; König et al., 2018). Analogously, research findings show that unemployed individuals show a low uptake of digital media (Witte et al., 2013; Helsper and Reisdorf, 2017). Thus, both age and unemployment are shown to increase the risk of entering into a "digital underclass" (Helsper, 2011; Helsper and Reisdorf, 2017) Therefore, it may be assumed, with initial research supporting this assumption (Feuls et al., 2016), that people falling into both categories, i.e. older unemployed individuals, are especially at risk of being further marginalized through digitalisation of public services.

At the same time, unemployment among older people is becoming an increasingly pressing societal issue. Due to the demographic changes across the world, the share of older people increases (United Nations, 2019b). However, older people have a smaller chance of reintegrating in the job market when losing their job (Vansteenkiste et al., 2015; Wanberg et al., 2016). Research suggests that digital media may help to improve their situation. In fact, studies show positive effects of ICT on the lives of older people (Morris et al., 2007; Charness and Boot, 2009; Ordonez et al., 2011; Cotten et al., 2012; Hill et al., 2015; Delello and McWhorter, 2017; Lüders and Brandtzæg, 2017) and of unemployed individuals (Garg and Telang, 2012, 2017; Suphan et al., 2012; Kuhn and Mansour, 2014; Felgenhauer et al., 2019; Klier et al., 2019) which in turn indicates ICT's potential to improve the situation of older unemployed people as well. As a consequence, research and practice are required to enhance the exploitation of this potential.

One promising and established way to encounter this challenge is the identification of user typologies. It allows researchers to better understand and structure large amounts of users and allows practitioners to provide adequate applications for specific user groups (Barnes et al., 2007; Brandtzæg, 2010). Despite the vast literature providing user typologies (e.g., Brandtzæg, 2010; Blank and Groselj, 2014; Borg and Smith, 2018), to the best of our knowledge, no typology on older unemployed people has been developed so far.

Our study aims to fill this gap and answer the following research question: What are the different digital media user types of older unemployed individuals? To provide an answer to this question, we conduct a survey on older unemployed people in cooperation with the German Federal Employment Agency and apply cluster analysis to give more nuanced insights into people's usage behaviour of digital media services (for job search), their attitude with respect to those services, currently perceived barriers and product preferences. Our contribution to research and practice is threefold. First, we contribute to the literature on digital media user typologies and particularly, to the understanding of the older and unemployed. Second, we show that a majority of older unemployed people are accessible for digital media in the context of job search. Third, we show pathways of how our findings can be used to inform the design of novel digital public services in the context of older unemployed people.

The remainder of this paper is structured as follows: In the next section, we provide an overview on related research. Section 3 describes the data collection process, our dataset and the research method used. Section 4 presents our findings. In Section 5, we critically discuss implications and limitations of our study and provide directions of further research. Finally, we conclude with a brief summary.

2 Related Work

2.1 Digital media and (un)employed individuals

Digitalisation plays an important role for business and society (United Nations, 2019a). According to the United Nations (2019a, p. 4), "opportunities created by the application of digital technologies are paralleled by stark abuses and unintended consequences. Digital dividends co-exist with digital divides." In the context of unemployment, there are initial indications in research that various forms of digital media can have a positive impact on the situation of the unemployed. For instance, Kuhn and Mansour (2014) show that online job search is more successful than staying offline. Klier et al. (2019) and Felgenhauer et al. (2019) indicate that mobile services like online peer groups improve participants' job search self-efficacy, self-exploration, and environmental exploration which in turn helps them to become employed again. Also, strategic engagement in social networks can positively influence job search effort and outcomes (Garg and Telang, 2012, 2017). However, digital services are not only promising with respect to the chances of reemployment, but also help individuals to deal with their situation in a more personal way. Particularly amongst the older unemployed, participation in social networks contributes to an improved well-being (Suphan et al., 2012). Also, regarding public services in the context of unemployment, digital services are a promising means. For instance, smart city technologies have the potential to improve job markets by installing e-career centres, creating regional hiring platforms and building data-driven retraining programs (Woetzel et al., 2018). Further, public services which are

facilitated by digital media improve the supply of labour market information and allow for a responsive and interactive service model, without the restrictions of opening hours or specific location (Pope, 2003; McQuaid et al., 2004; Lindsay, 2005).

Looking behind the scenes, research indicates that one possible reason for the promising benefits of digital services might be their online communication features: Accessibility, disinhibition, and written interaction mode were shown to elevate the effects of career counselling services for unemployed individuals (Klier et al., 2019). Accessibility renders time and location irrelevant, allowing for support outside working hours and without relying solely on local specialists (White, 2001; Cook and Doyle, 2002; Coulson, 2005). Moreover, disinhibition through anonymous communication fosters self-disclosure as e.g., fear of embarrassment and judgment declines (Cook and Doyle, 2002; Amichai-Hamburger and Furnham, 2007; Wildevuur and Simonse, 2015). Lastly, a written interaction mode allows for better cognitive processing, with the opportunity to reflect and facilitation of expressing thoughts and feelings (Cook and Doyle, 2002; Coulson, 2005).

However, despite the wide range of potential benefits digital services offer for unemployed people, current literature suggests that this potential cannot yet be fully exploited. This is because, in contrast to the omnipresence of digital services among wide parts of the society (Yoo, 2010), unemployed individuals lack behind in the usage of digital services (Witte et al., 2013; Helsper and Reisdorf, 2017). One factor that further hampers the usage of digital services amongst unemployed, is rising age (Feuls et al., 2016). Therefore, our study focuses on the most vulnerable group of unemployed people with respect to digital media usage: older unemployed people.

2.2 Digital media and older workforce

Among unemployed people, chances for finding a job vary depending on different factors. One of those is age, with older unemployed people over age 50 having a reduced chance of reintegrating into the labour market (Vansteenkiste et al., 2015; Wanberg et al., 2016). Within the scope of our study, we focus on unemployed individuals aged between 50 and the pensionable age, i.e. the older workforce (Nolan and Barrett, 2018; Vettori, 2016; Moen et al., 2017; Kumar and Srivastava, 2018; Sun et al., 2020). As the share of older people increases in the world's population (United Nations, 2019b), unemployment amongst older individuals becomes more and more relevant for societies. Consequently, there is a need for effective support services for older unemployed individuals. To the best of our knowledge, there is a lack of research on exactly this target group so far. However, research on older people might give first indications of digital media use among older unemployed individuals.

Digital media has the potential to positively impact the lives of older individuals. In fact, research shows that digital media usage can improve cognitive abilities through the learning of new technologies (Charness and Boot, 2009; Ordonez et al., 2011), foster self-empowerment (Hill et al., 2015), positively influence the connection with friends and family (Morris et al., 2007; Hill et al., 2015; Lüders and Brandtzæg, 2017), serve a general desire to stay up to date (Morris et al., 2007) and improve mental well-being (Cotten et al., 2012).

However, older individuals still remain overrepresented amongst the digitally excluded (Hunsaker and Hargittai, 2018; König et al., 2018). Analogously to the slow adoption of ICT in general, older individuals show a low usage rate of digital public services (Niehaves and Becker, 2008; United Nations, 2018). The low uptake of digital services is specifically problematic as on the one hand, these services do not reach their full potential to serve older citizens and on the other hand contributes to further exclusion as more services move online (United Nations, 2018).

This lack of usage attracted wide research attention, with lots of studies focusing on identifying barriers with respect to ICT usage amongst older users. These range from structural issues such as a lack of Internet access (Morris et al., 2007; Yu et al., 2016; König et al., 2018) and (the perception of) excessive usage costs (Lee et al., 2011), over causes related to skills and experiences such as a lack of knowledge around basic functionalities, e.g. how to access the Internet (Olphert et al., 2005; Yu et al., 2016; Lüders and Brandtzæg, 2017), to causes related to a lack of perceived utility and an overall negative attitude towards digital media (Olphert et al., 2005; Morris et al., 2007; Lüders and Brandtzæg, 2017) up to

privacy and security concerns (Olphert et al., 2005; Lüders and Brandtzæg, 2017). Besides, research identifies physical, economic or sociocultural disadvantages as further impediments among older individuals regarding digital media (Morris et al., 2007; Lee et al., 2011; Vošner et al., 2016; Yu et al., 2016). Therefore, to better serve older digital media users, it is necessary to differentiate by context, rather than adressing older users as one homogenous group. This is exactly what our study aims to address by sheding light on digital media usage amongst the older unemployed.

2.3 Digital media user typologies

One method to make variations of user participation and behaviour more tangible, to gain a better understanding on numerous individuals and to provide a basis for the development of valuable services for specific user types is to extract user typologies, thus distinguishing user types based on varying usage patterns (Barnes et al., 2007; Brandtzæg, 2010).

Interest in better understanding user behaviour with regards to digital media has generated a vast amount of literature (e.g., Brandtzæg, 2010; Blank and Groselj, 2014; Borg and Smith, 2018), with some studies focusing on specific applications such as social networks (e.g., Brandtzæg and Heim, 2011; Brandtzaeg, 2012; Akar et al., 2019; Fortier and Burkell, 2018; Kim, 2018) or specific user groups such as students (Zawacki-Richter et al., 2015) or unemployed (Feuls et al., 2016). One of the most prominent typologies is the one derived by Brandtzæg (2010). It builds upon literature on the use of so-called "new media" (i.e. "television, computers, Internet, different game consoles, mobile phones"), the Internet and different Internet services like for example social networking sites (Brandtzæg, 2010, p. 943) and can also be well mapped to typologies in later studies, even though the naming of the categories sometimes varies. The two most extreme user types are *Non-users* who do not make use of the digital media considered in the respective study on the one side (e.g., Selwyn et al., 2005; Ortega Egea et al., 2007; Brandtzæg, 2010; Feuls et al., 2016; Borg and Smith, 2018), and *Advanced users* who very frequently use a large variety of digital media applications on the other side (e.g., Brandtzæg et al., 2005; Brandtzæg, 2010; Distel and Becker, 2017). In between, Brandtzæg (2010) identifies *Sporadics* (e.g., Kau et al., 2003; Ortega Egea et al., 2007; Brandtzæg, 2010; Brandtzæg and Heim, 2011) and *Lurkers* (e.g., Kau et al., 2003; Brandtzæg, 2010; Brandtzæg and Heim, 2011; Akar et al., 2019) who do use digital media, but in a very limited way, with *Sporadics* being characterised by a very low usage frequency and variety and *Lurkers* passively consuming digital media, often to "kill some time" (Brandtzæg and Heim, 2011, p.41). Finally, there are four types of users showing a medium frequency and variety of new media usage. *Debaters* mainly use blogs and social networking sites and actively contribute to these media (e.g., Brandtzæg, 2010; Brandtzæg and Heim, 2011), *Socializers* are mainly active on social networking sites to build and maintain social relationships (e.g., Johnson and Kulpa, 2007; Brandtzæg, 2010; Füller et al., 2014), *Instrumental users* pursue utility and information goals when using new media (e.g., Howard et al., 2001; Johnson and Kulpa, 2007; Brandtzæg, 2010; Borg and Smith, 2018) and finally, *Entertainment users* mainly use new media for entertainment purposes (e.g., Heim et al., 2007; Brandtzæg, 2010).

There is a range of studies focusing on particular applications, such as social networks (e.g., Brandtzæg and Heim, 2011; Brandtzaeg, 2012; Akar et al., 2019; Fortier and Burkell, 2018; Kim, 2018). For instance, Kim (2018) analyses different perceptional typologies with respect to social network usage and identifies the four types *Impression Management Type*, *Lurker Type*, *SNS Enjoyer & Relationship Focus Type* and *Social Value Orientation Type*. Further, there is a range of studies focusing on specific groups of people. For instance, Zawacki-Richter et al. (2015) study media usage patterns among students and Brandtzæg et al. (2005) study Internet use among children. The group of unemployed people is addressed in Feuls et al. (2016) who determine user types based on how unemployed use social media and differences in terms of access, skills and motives. They figure out four types of users: *Non-users*, *Novices* who lack experience lack of experience and skills with the Internet and show concerns, *Passive users* who use the Internet on a regular base, and *Heavy users* who can be mapped to *Advanced users* as defined above. Despite the wide range of literature providing typologies of user behaviour, to the best of our knowledge no specific classification exists regarding the older fraction of the group of unemployed people.

2.4 Research question and contribution of this study

Existing literature highlights that older unemployed individuals face the double burden of age and unemployment (Feuls et al., 2016; Yu et al., 2016). Even though studies show that both aspects go along with a reduced usage of digital services (Witte et al., 2013; Helsper and Reisdorf, 2017; Hunsaker and Hargittai, 2018; König et al., 2018), research also indicates a large potential of digital services to improve the situation of both unemployed (Garg and Telang, 2012, 2017; Suphan et al., 2012; Kuhn and Mansour, 2014; Felgenhauer et al., 2019; Klier et al., 2019) and older people (Morris et al., 2007; Charness and Boot, 2009; Ordonez et al., 2011; Cotten et al., 2012; Hill et al., 2015; Delello and McWhorter, 2017; Lüders and Brandtzæg, 2017) and thus presumably also of older unemployed people. Consequently, there is a need for studies trying to better understand this target group in order to address problems of digital divide and conceptualize suitable digital services for people in this target group. Classifying the respective target group into subgroups according to usage is one approach to achieve this (Brandtzæg, 2010; Brandtzæg et al., 2011). In our study, we take this approach to answer the following research question:

RQ: What are the different digital media user types of older unemployed individuals?

In cooperation with the German Federal Employment Agency, we conduct a survey on older unemployed people and apply cluster analysis to give more nuanced insights into people's usage of digital media services (for job search), their attitude with respect to those services, currently perceived barriers and product preferences. Our contribution to research and practice is threefold.

First, by providing a user typology we contribute to the understanding of older unemployed individuals as a nuanced group with a great range of different usage behaviours and attitudes towards digital media (for job search). Thus, we provide a theoretical contribution by permitting exploration into the sources and consequences of different user types amongst older unemployed and provide another media usage pattern for comparisons with other studies on media usage among specific user groups. Second, we show that contrary to expectations based on literature, a large group of older unemployed individuals is accessible for digital media in the context of job search. Thereby, from a theoretical point of view, our study updates insights from prior literature characterising both unemployed and individuals above 50 as consisting of many "Non-users" and provides an initial starting point for research to further investigate digital media usage amongst marginalized groups. Further, we provide a practical contribution of exhibiting the large potential for digital services for that user group. Third, our user typology containing information about digital media usage (for job search), attitude towards digital media for job search and perceived barriers to this respect, allows us to identify pathways of how digital media usage among older unemployed people can be further improved to be able to exploit its potential. This way, we provide a practical contribution to assist product developers to create services that better fit this target group and practitioners who aim to promote services to that group (Brandtzæg, 2010).

3 Research Methodology

In this section, we introduce the case setting, our data collection, the resulting dataset and finally, we describe our data analysis process.

3.1 Case setting and data collection

This study was made possible thanks to a cooperation with the German Federal Employment Agency (Bundesagentur für Arbeit). With 156 local agencies, about 600 branch offices and more than 95,000 employees, the Federal Employment Agency is the main supplier of labour market services in Germany (Federal Employment Agency, 2018). It is mainly known for its services for private citizens, like career counselling, employment placement, vocational guidance and financial support.

Within the scope of our analyses, we define older unemployed people as unemployed people of age 50 or older. We decided to use questionnaire-based surveys to derive a user typology of older unemployed people in our study as this is the most common means employed by studies focusing on classifying

digital media users into user types (cf. Brandtzæg, 2010). Both our survey setting, and the design of our questionnaire were realised under consideration of remedies to reduce response bias (e.g., Podsakoff et al., 2012; Menold and Bogner, 2016). The survey is structured as follows: First, we assess current usage by identifying the range and frequency of private usage (cf. Feuls et al., 2016) and usage and obstacles in the job search context (cf. Kuhn and Mansour, 2014; Feuls et al., 2016; Garg and Telang, 2017). Second, we use a Likert-type scale ranging from 1 (*"applies not at all"*) to 6 (*"fully applies"*) to enquire whether people think digital media helps them in their job search, whether they wish for more digital services in the context of job search and to what extent they would like future products to feature different aspects. We used two control variables: gender and age (cf. Morris et al., 2007; Lee et al., 2011; Yu et al., 2016).[1]

The survey was distributed amongst all 19 Employment Agencies in the German state Baden-Württemberg, the third-largest state in Germany. This wide spreading allowed us to both study urban and rural regions and to potentially reach every unemployed citizen in Baden-Württemberg who fulfilled our target group criteria. The survey was provided to the Employment Agencies in hard copy, to avoid attracting more digital-savvy participants by providing a digital format. The Employment Agencies were then asked to briefly introduce the survey to all customers within the relevant age group, i.e. 50 years or older, at the end of each counselling session and to offer them participation in our survey on a voluntary base. Surveys on which information on the participant's age or on the clustering variable *How often do you use digital media in everyday life?* was missing and surveys indicating that the participant was younger than 50 were excluded from the study resulting in 192 valid surveys.

3.2 Dataset

The dataset at hand contains the data from the 192 valid surveys. There is a slightly higher share of males than females among the associated participants while the group of people above 60 is slightly smaller than the other two groups. Table 1 provides an overview on the demographics in the survey. The upper limit of 65 years can be explained by the pensionable age in Germany which is currently at 65 years and nine months for people born in 1955 (Deutsche Rentenversicherung, 2019).

	Number of participants (without NAs)	Male participants	Female participants	Participants age ≤55	Participants age >55, ≤60	Participants age >60, ≤65
Age data	192	105	86	74	76	42

Table 1. Distribution of age in the survey.

3.3 Data analysis

Our analysis aims at identifying the potential of digital media usage in the context of job search and at deriving strategies for a better adoption of digital media applications for job search.

We first explore different digital media user types by applying two-step clustering in SPSS on the variables *Which of the following digital media communication applications do you use?* and *How often do you use digital media in everyday life?*. We chose this method as it does not require to choose the number of clusters beforehand but determines it through optimisation (IBM, 2019). In our case, optimisation was realised according to the Bayesian Information Criterion (BIC). It results in four clusters with a positive average silhouette score (*BIC*=1123.339, *silhouette*=0.577), suggesting a strong clustering structure with an interpretable number of clusters: *Digital Sceptics*, *Digital Classics*, *Digital Interactives* and *Digital Contributors*. A silhouette plot, as shown in Figure 1, helps to assess the quality of the clusters (Rousseeuw, 1987). We find that the clear majority of observations have a good fit within their

[1] The survey can be accessed online via the following link: https://www.uni-ulm.de/fileadmin/website_uni_ulm/mawi.inst.125/OnlineAppendix_SurveyEnglish.pdf

clusters. However, the cluster referred to as *Digital Sceptics* seems to contain several observations that lie between clusters.

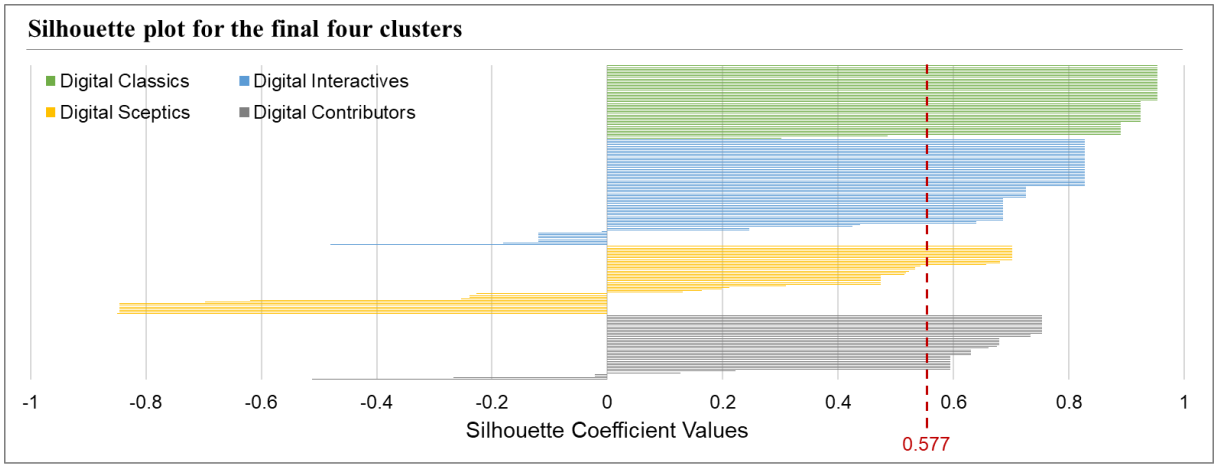


Figure 1: Silhouette plot for the final clustering result.

In a second step, we use descriptive analyses on the categorical data as well as one-sample t-tests with μ=3.5 for the variables with Likert-type scale to characterise the clusters. Finally, we perform comparative analyses across the clusters. Here, we use Chi-Squared tests (i.e. the R function chisq.test() from base R (R Core Team, 2018)) for group comparisons with respect to sex and with respect to the categorical variable *In case there are digital media you do not use for job search: What is the main reason for that?*. We apply Mann Whitney tests for group comparisons with respect to age and with respect to items with Likert-type scale (*Digital media assist me in my job search, I wish for more digital offers to assist me in my job search, I regularly receive relevant information regarding job search, I can ask all of my questions and get answers, I can learn from other job-seekers' experiences and share my own experiences*). Precisely, we use the function *wilcox.exact()* from the R Package *exactRankTests* () which applies the Shift-Algorithm by (Streitberg and Röhmel, 1986). Concerning the last three items, we conduct Wilcoxon signed rank tests (R function *wilcox.test()* from base R (R Core Team, 2018)) to compare which of the items are considered most important across all participants.

4 Research Findings

This section is dedicated to the results of our study. We first present the user types identified by means of cluster analysis. This way we provide insights into older unemployed people's usage behaviour, attitudes, wishes, and barriers concerning digital services for job search. Second, we apply comparative statistics to work out similarities and differences relevant for theory and practice.

4.1 Results of cluster analysis

The cluster analysis based on the usage behaviour of the participants results in four clusters (*BIC*=1123.339, *silhouette*=0.577) which can be briefly characterised as *Digital Sceptics*, *Digital Classics*, *Digital Interactives*, and *Digital Contributors* (cf. Figure 2).

Digital Sceptics are people with an overall low usage of digital media and an overall negative attitude towards the potential for digital media for job search. The group contains of 22% of all participants and consists of about 60% male and 40% female participants, with an average age of 56.83 years. We find that 26% of *Digital Sceptics* use digital media daily, with another 26% of participants once a week and 29% of participants reporting to never use digital media in everyday life. Half of the *Digital Sceptics* (45%) use exactly one of the given digital media applications. We identify e-mail and messaging services as the most popular applications in this group, with almost no *Digital Sceptic* reporting to participate in Internet forums or publish own contents. We find that 55% of *Digital Sceptics* do not use digital

media for job search. The main problems of digital media use in the context of job search in this group are a lack of dealing with digital media applications for job search, as stated by 32% of *Digital Sceptics*, and concerns with respect to data security, as stated by 29% of individuals in this group. We identify that *Digital Sceptics* are not convinced that digital media assist them in job search (*mean*=2.74, *sd*=1.87; *p*=0.003) and do not wish for a larger supply of respective applications (*mean*=2.26, *sd*=1.61; *p*=0.000). Finally, there is no consent within the group concerning the importance of new digital offerings to provide regularly delivered, relevant information (*mean*=3.55, *sd*=2.00; *p*=0.872) or individual information (*mean*=3.71, *sd*=2.18, *p*=0.586). Finally, the group is rather uninterested in digital media allowing to exchange information with other job-seeking people (*mean*=2.91, *sd*=1.87, *p*=0.075).

Digital Classics are individuals that are rather conservative in their digital media usage having the highest average age amongst the four groups (57.60 years) and containing 23% of all participants. Interestingly, a large share of *Digital Classics* is male, with only 38% of *Digital Classics* being female. They use digital media rather frequently in everyday life with 49% of *Digital Classics* using digital media daily and only 18% less frequently than once in a month. We find that all *Digital Classics* use e-mail and less than 5% an additional application (e.g., video calls). Concerning job search, only 20% of the individuals in this group report to not use any digital media for this purpose. Beyond that, we find that the most popular application, employed by 58% of *Digital Classics* who use digital media for job search, are websites of companies, followed by job platforms, employed by 56% of them. We find that insufficient ease of use is the main barrier hindering the usage of digital media for job search within this group, stated by 22% of participants. Further, we find that 24% indicate to not have any barriers towards using digital media for job search. We identify no consent in this group concerning the assessment that digital media assists them in job search (*mean*=3.84, *sd*=1.76, *p*=0.558) and regarding their wish for a larger supply of according applications (*mean*=3.36, *sd*=1.76, *p*=0.607). Finally, the group appreciates regularly delivered, relevant information (*mean*=4.36, *sd*=1.65, *p*=0.002) and individual information (*mean*=4.46, *sd*=1.58, *p*=0.000), but is undecided on the importance of options for information exchange with other job-seeking people (*mean*=3.78, *sd*=1.78, *p*=0.320).

Digital Interactives are the largest of the four groups, including 34% of all participants in the survey. They use a wide range of digital media applications, which they employ frequently in their everyday life and perceive digital media as useful for their job search endeavour. The group of *Digital Interactives* consists of 55% females and 45% males, with an average age of 56.82 years. Everybody in this group uses digital media frequently in everyday life with 83% of the people indicating even a daily use. Concerning the communication patterns, again all people indicate using e-mail. However, 86% of the people additionally use messaging services, 40% social networks and 9% take part in an Internet forum. Concerning job search, less than 10% of the people in this group report to not use any digital media in this respect. Beyond that, the most frequently used applications in this context are job platforms (72%), websites of companies (68%) and search engines (60%). What is new within this group is the use of social networks (14%). The actual indications of main hurdles are rather equally distributed with the main hurdle being the fact that people have not yet engaged themselves with the topic (22%), followed by concerns with respect to data security (16%). We find that about 29% of *Digital Interactives* do not have barriers hindering digital media usage in the context of job search. Further, we can show that this group is convinced that digital media helps them in job search (*mean*=4.23, *sd*=1.78, p=0.001) but is still undecided with respect to a larger supply of respective applications (*mean*=3.76, *sd*=1.75, *p*=0.265). Finally, we identify that this group also appreciates digital products to provide regularly delivered, relevant information (*mean*=4.54, *sd*=1.71, *p*=0.000) and individual information (*mean*=4.60, *sd*=1.75, *p*=0.000), but is undecided on the importance of options for information exchange with other job-seeking people (*mean*=3.57, *sd*=1.69, *p*=0.734).

Digital Contributors are the most progressive users with almost everybody in this group using digital media on a daily base (98%). This group consisting of 21% of participants, 43% of them being female and 57% male, has an average age of 55.38 years. Overall, this group shows high digital media activity and is accessible via digital communication: The most prominent digital application is e-mail (100%), followed by messaging (93%) and video calls (93%). What is new in this group is the predominant usage of video calls (93%) as well as the high fraction of participants in Internet forums (38%) or publishers

of own contents (25%). Therefore, people in this group are not only consumers of digital services and products, but actively contribute through sharing own content to the digital ecosystem. Concerning job search, 95% report to use digital media in this respect. Beyond that, the most frequently used applications in this context are job platforms (93%), websites of companies (88%) and career networks (75%). We find that the main hurdles show a peak for data security concerns, stated by 26% of *Digital Contributors*, whereas almost nobody misses a respective benefit (3%). This group is convinced that digital media helps them in job search (*mean*=5.10, *sd*=1.13, *p*=0.000) and shows a desire for more respective applications (*mean*=4.36, *sd*=1.50, *p*=0.001). Finally, we identify that *Digital Contributors* would appreciate future digital products to provide regularly delivered, relevant information (*mean*=4.97, *sd*=1.24, *p*=0.000) and individual information (*mean*=5.13, *sd*=1.28, *p*=0.000), but are undecided on the importance of options for information exchange with other job-seeking people (*mean*=3.74, *sd*=1.74, *p*=0.406). Figure 2 summarizes the usage behaviour in the four groups and according demographic data.

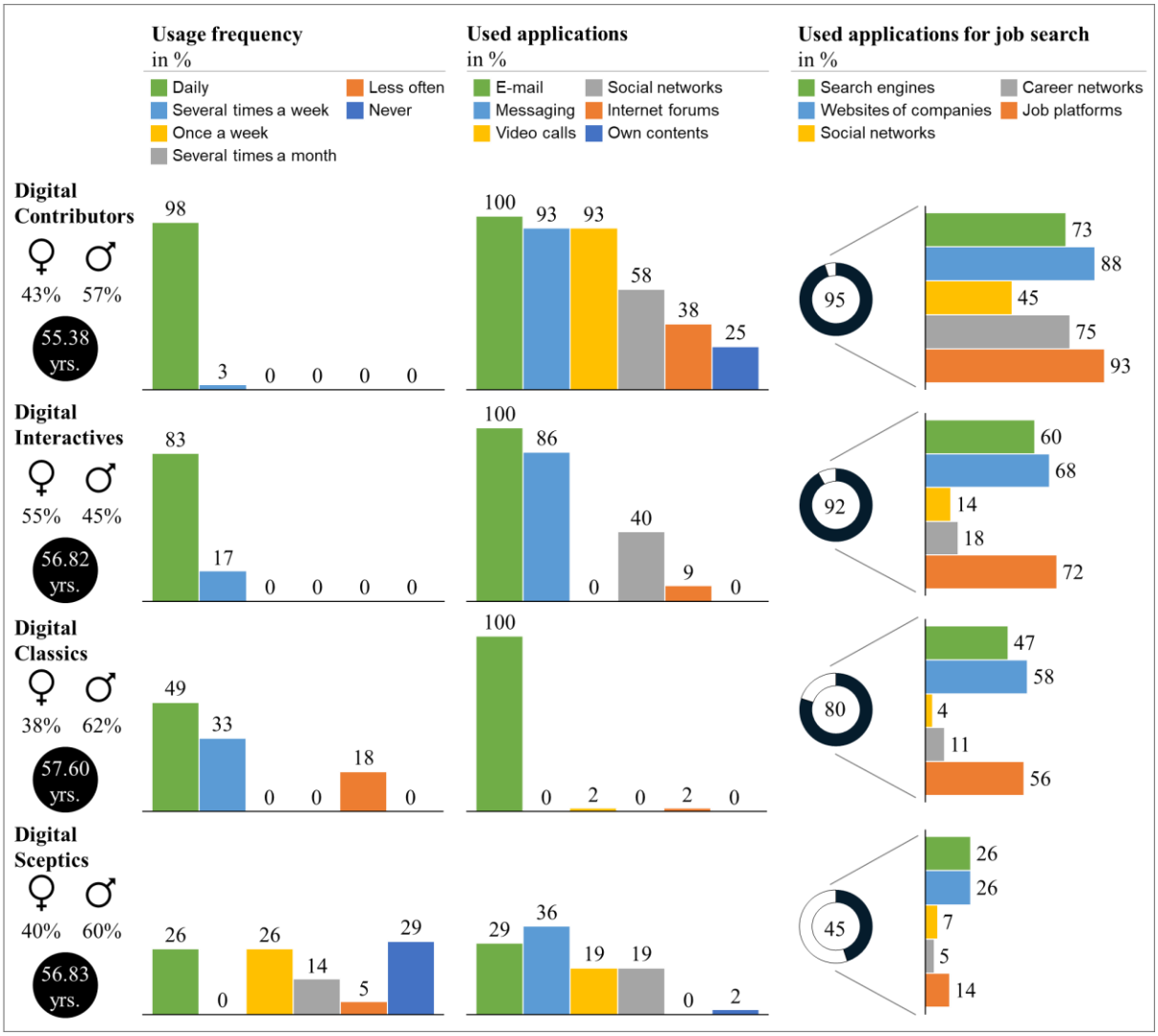


Figure 2. Summary of demographic data and usage behaviour of the four clusters.

4.2 Results of comparative analyses

Apart from providing an overview on the characteristics of the four user types, our study sheds light on similarities and differences across groups based on comparative statistics. Particularly, we identify patterns going along with changes in the intensity and variety of ICT usage.

First, we can show that neither age nor gender play a relevant role in distinguishing the groups from one another with respect to digital media usage. A chi-squared test does not reveal significant differences between the shares of males and females across the four groups (p=0.435). Analogously, two-sample Wilcoxon tests on the combinations of close-by groups with respect to digital media usage do not reveal significant differences with respect to age (p(*Digital Sceptics*, *Digital Classics*)=0.407, p(*Digital Classics*, *Digital Interactives*)=0.306, p(*Digital Interactives*, *Digital Contributors*)=0.068).

Second, both people's assessment that digital media assist them in job search (1) and their wish for more digital offerings for job search (2) increase across the groups. Particularly, this increase runs parallel to the increase in the frequency and variety of digital media use across the groups, i.e. from *Digital Sceptics* to *Digital Classics* to *Digital Interactives* to finally, *Digital Contributors*. Significant differences between adjacent groups can be observed at the margins (p_1(*Digital Sceptics*, *Digital Classics*)=0.003; p_1(*Digital Interactives*, *Digital Contributors*)=0.011; p_2(*Digital Sceptics*, *Digital Classics*)=0.002; p_2(*Digital Interactives*, *Digital Contributors*)=0.010). Figure 3 illustrates these insights.

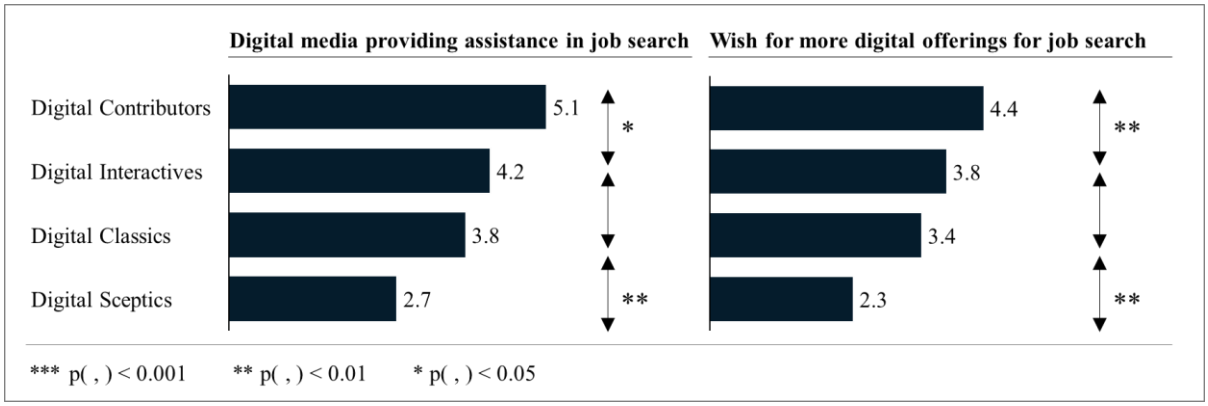


Figure 3. Assessment of digital media assisting job search and wish for more digital offerings for job search.

Third, our analysis shows that all four user groups have similar preferences regarding specific aspects future digital products for job search should provide. The two rather conservative aspects, i.e. receiving regular, relevant information regarding job search and the possibility to ask own questions and get answers, are valued more important than the opportunity to learn from other job-seekers' experiences and share own experiences. Particularly, (one-sided) paired-samples Wilcoxon tests show that there is no significant difference in the users' valuation of the two conservative aspects (p=0.097) whereas the depicted importance of opportunities of experience exchange among job-seeking people is significantly lower than that of regular, relevant information (p=0.000).

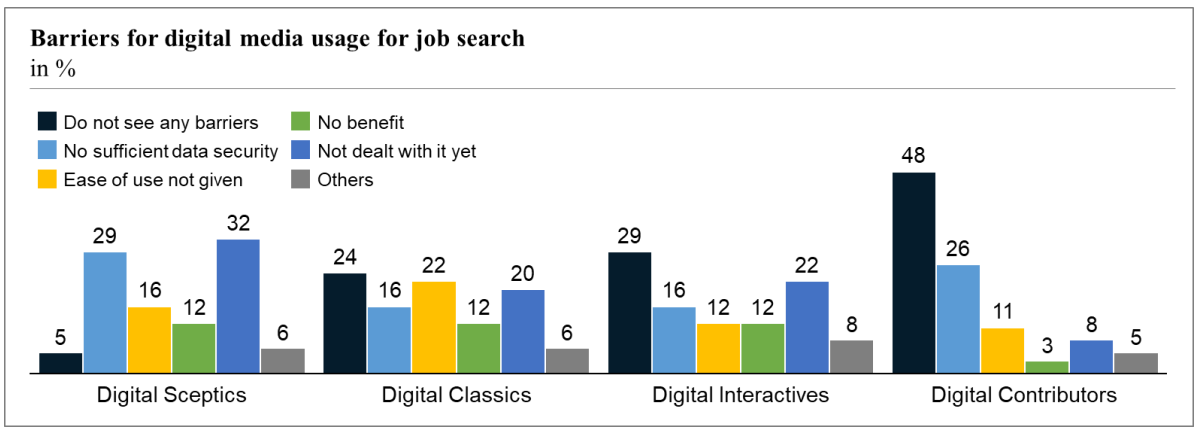


Figure 4. Barriers for digital media usage for job search in %.

Finally, two main hurdles with respect to digital media usage in the context of job search deserve closer attention. First, a high share of those *Digital Contributors* who indicate barriers with respect to digital media usage for job search, see the main problem in data security issues. Second, among *Digital Classics*, it is mainly the deficiency in the ease of use which keeps them from using certain digital media applications for job search. Figure 4 shows the distribution of barriers across the groups.

5 Discussion, Limitations, and Future Research

5.1 Implications for theory and practice

Our research aimed to better understand individuals facing the double burden of age and unemployment, oftentimes perceived as a "digital underclass" (Helsper, 2011; Helsper and Reisdorf, 2017), with respect to their digital media usage (for job search). For this purpose, we conducted a survey on older unemployed people and applied cluster analysis to give more nuanced insights into people's usage behaviour of digital media services (for job search), their attitude with respect to those services, currently perceived barriers and product preferences. This way we complement literature on digital media user typologies. Our findings lead to three main contributions for theory and practice.

First, we demonstrate that older unemployed individuals are indeed far away from being a homogenous group with respect to usage behaviour of digital media services (for job search). Instead, we show that older unemployed people can be grouped into four different types of digital media users – *Digital Sceptics, Digital Classics, Digital Interactives*, and *Digital Contributors* – which exploit a wide range of different usage behaviours and attitudes towards digital media for job search. While on the one hand these user types are to certain extent similar to those identified by prior research (Brandtzæg, 2010; Brandtzæg et al., 2011; Brandtzæg and Heim, 2011; Brandtzaeg, 2012; Feuls et al., 2016; Akar et al., 2019), we also observe surprising differences. Firstly, most prior literature identifies a user group, often called "Non-users", which is characterized by a lack of digital media usage (c.f., Brandtzæg, 2010; Brandtzæg et al., 2011; Feuls et al., 2016). Feuls et al., (2016) further show that among the unemployed people in their study, "Non-users" are predominantly older than 50 years, and thus expect a larger share of "Non-users" among this age group. Thus, it is surprising to find that within our sample of older unemployed individuals, the portion of people indicating to never use digital media is rather small and does not compose a single user type but falls in the group of *Digital Sceptics*. *Digital Sceptics* are mainly characterised through a negative attitude towards digital media usage for job search and compared to literature, reflect the user type "Sporadics", similar in the low overall use (Brandtzæg, 2010; Brandtzæg et al., 2011; Brandtzæg and Heim, 2011; Brandtzaeg, 2012) or "Novices" similar with respect to their lack of experience and skills and their concerns (Feuls et al., 2016). The second group we identify are *Digital Classics* which resemble "Instrumental Users" regarding the focus on specific digital activities (Brandtzæg, 2010; Brandtzæg et al., 2011). The third group identified in our study, *Digital Interactives*, can be matched to the user type "Socializers" identified in former studies (Brandtzæg, 2010; Brandtzaeg, 2012) as both use messaging services, potentially to interact with friends, family and new contacts. Lastly, rather than identifying "Lurkers" (Brandtzæg, 2010; Brandtzæg and Heim, 2011; Brandtzaeg, 2012; Kim, 2018) and "Passive users" (Feuls et al., 2016), our study results raise up the converse and identify *Digital Contributors* as individuals who actively shape the digital discussion by contributing own contents, similar to "Content Generators" (Akar et al., 2019). In turn, we also do not specify a single group as "Advanced Users" characterized in literature through their high frequency of and variety in (social) media usage (Brandtzaeg, 2012), as all clusters in our study, except *Digital Sceptics*, exhibit a high frequency of use, with the two clusters *Digital Interactives* and *Digital Contributors* also showing a high variety of use. Thus, in our study, we could not observe frequency and variety as sufficient to distinguish the usage clusters. Hence, by taking a broader perspective and also including attitude towards digital media in our observation, we offer a nuanced insight into the group of unemployed older people. Thereby, we provide a theoretical contribution by permitting exploration into the sources and consequences of different user types amongst older unemployed and provide another media usage pattern for comparisons with other studies on media usage among specific user groups. From a practical point of

view, the four clusters can serve as an orientation in addressing older unemployed individuals in the context of digital media.

Second, against conventional wisdom that older unemployed individuals are seen as "digital underclass" (Helsper, 2011; Helsper and Reisdorf, 2017) and being particularly affected by digital divide (Feuls et al., 2016; Yu et al., 2016), we show that three out of four clusters – *Digital Classics*, *Digital Interactives*, and *Digital Contributors* – are accessible for digital media in job search. This is also surprising given the literature on ICT usage behaviour of older people and unemployed individuals which suggests that the double burden of age and unemployment contributes to this group being particularly affected by digital divide (Feuls et al., 2016; Yu et al., 2016; Helsper and Reisdorf, 2017). In particular, prior typologies show that within the group of "Non-users", older individuals are overrepresented (Brandtzæg et al., 2011; Feuls et al., 2016), while "Advanced users" and "Heavy users" mainly consist of younger individuals. Against this backdrop, it is striking that within our sample of older unemployed, "Non-users" do not even compose an own group, whereas *Digital Interactives*, a group characterized by high frequency and variety of usage, is the largest group observed and we even observe *Digital Contributors* who actively shape the digital space. Thus, we show that the older unemployed are willing to participate in digital services increasing their chance for reemployment (Garg and Telang, 2012, 2017; Kuhn and Mansour, 2014), and gaining a chance for improved well-being (Suphan et al., 2012). Thereby, our results update insights from prior research characterising both unemployed and individuals above 50 as consisting of many "Non-users" and might encourage researchers to further investigate digital media usage amongst other marginalized groups.

Third, our study sheds light on how to further support digital media use in job search among this share of accessible older unemployed people. A large cluster which is situated rather at the beginning of the usage scale consists of *Digital Classics*. Today, their usage is restricted to very established applications such as e-mail which is used by everybody in this cluster. Shifting this group bears a large potential in leading older unemployed people to a more advanced usage of digital media for job search, for example using social networking applications or sharing own contents. This is particularly beneficial because more recent, advanced types of services, like for example mobile peer groups, have been shown to improve chances of reintegration into the job market (Klier et al. 2019). To date, the major barrier for use of digital media in job search among *Digital Classics*, ease of usage, should be considered in the design of services. To achieve a high degree of usability, an approach based on human-centred design as central element is needed, incorporating insights on the specific needs of older users (e.g., Boll and Brune, 2015; Nurgalieva et al., 2019). Increasing people's digital skills further represents another possible step in shifting this group to a higher usage of digital media for job search. Even though situated at the upper end of the usage scale, *Digital Contributors* can be further supported in their use of digital media for job search. Based on their indications of current barriers, especially a high level of data security as well as a high level of transparency with respect to this issue are inevitable when addressing this group.

5.2 Limitations and future research

Although our findings provide first and interesting insights into the usage behaviour of older unemployed individuals, our study has several limitations which could be addressed in future studies. First of all, our investigation is a first step with a limited number of participants in one country and specifically one state (N=192). As culture, job market situations, and the level of digitalisation in public services vary among countries and there might also be slight differences across one country, our findings may only be generalisable to certain extent. Second, the scope of the survey and thus also the number of variables considered in this study is limited due to restrictions from the Federal Employment Agency. Nevertheless, the close cooperation with the Federal Employment Agency has offered us the unique opportunity to gain insights into a sensitive problem context, that up to now received little research attention. Third, our sample may evidence a self-selection bias: participants who engaged in our survey might be people who have an affinity to digital media. To address this bias, we selected participants randomly to fill out the survey in the agencies.

We suggest that future research should address these limitations, further generalise our results and extend our insights to further analyses. First, by repeating surveys with a larger number of participants in other states of Germany and even other countries, researchers might confirm our analysis. A larger dataset would further allow for further differentiations with respect to each participant's characteristics. Against the backdrop of demographic developments omnipresent globally (United Nations, 2019b), and the lower chance for reintegration into the job market for older unemployed people (Vansteenkiste et al., 2015; Wanberg et al., 2016), the topic is of rising relevance. Second, within the scope of larger future studies, we suggest researchers to rerun the two-step clustering and additionally further clustering algorithms to see whether the identification of *Digital Sceptics* is robust. If so, studies on this user group might further investigate the underlying causes of their negative attitude towards digital media (for job search) and further characteristics of this group. Third, research findings might benefit from the extension of further variables. Adding demographic variables such as level of education, allows to study whether different demographic characteristics are predominant in the different clusters. Further, we suggest that future research could provide a differentiated perspective on different fields of work. As for example online job postings and the need for certain digital competencies differ by various fields of work, it would be useful to provide a nuanced understanding of the user behaviour of the older workforce by various professional clusters. Finally, future studies could analyse which features of digital media services for job search older unemployed individuals assess the most important and beneficial. For example, first studies in the context of unemployment show that mobile peer groups improve the level of information, emotional support, comfort, and support of the participants overall (Felgenhauer et al., 2019; Klier et al., 2019).

6 Conclusion

While ICT are increasingly present in public services globally (United Nations, 2018), marginalized groups show lower usage rates of such services (Niehaves and Becker, 2008; United Nations, 2018). Both older (Hunsaker and Hargittai, 2018; König et al., 2018) and unemployed individuals (Witte et al., 2013; Helsper and Reisdorf, 2017) are affected by this risk of entering into a "digital underclass" (Helsper, 2011; Helsper and Reisdorf, 2017). Thus, it may be assumed, with research supporting this assumption (Feuls et al., 2016), that older unemployed people are especially affected by this risk. At the same time, literature suggests that digital media may improve the situation of both the older (Morris et al., 2007; Charness and Boot, 2009; Ordonez et al., 2011; Cotten et al., 2012; Hill et al., 2015; Delello and McWhorter, 2017; Lüders and Brandtzæg, 2017) and the unemployed (Garg and Telang, 2012, 2017; Suphan et al., 2012; Kuhn and Mansour, 2014; Felgenhauer et al., 2019; Klier et al., 2019). These insights suggest digital media to have a potential in supporting older unemployed people as well. Consequently, both research and practice are required to enhance the understanding of user behaviour in this group to counter the phenomenon of digital divide among older unemployed and thus fully exploit the potential of ICT. One approach to achieve this goal is to identify user typologies in the respective target group (Brandtzæg, 2010). Against this backdrop, this study aimed to answer the following research question: What are the different digital media user types of older unemployed individuals? To provide an answer, in cooperation with the German Federal Employment Agency we conducted a survey on older unemployed people and applied cluster analysis to give more nuanced insights into people's usage behaviour of digital media services (for job search). Our findings suggest that older unemployed individuals are a nuanced group with a great range of different usage behaviours and attitudes towards digital media for job search. We further identify that a large group of older unemployed individuals is accessible for digital media in the context of job search. Finally, we shed light on how to further support digital media use in job search amongst the accessible older unemployed people. We believe that our study is a first but important step towards a better understanding of digital media usage in the context of the older unemployed. We hope that our results will stimulate further research on that fascinating topic and will serve as a useful starting point for future research as well as for the practical improvement of digital media usage among older unemployed individuals.

References

Akar, E., S. Mardikyan and T. Dalgic (2019). "User Roles in Online Communities and Their Moderating Effect on Online Community Usage Intention: An Integrated Approach." *International Journal of Human–Computer Interaction* 35 (6), 495–509.

Alshibly, H. and R. Chiong (2015). "Customer empowerment: Does it influence electronic government success? A citizen-centric perspective." *Electronic Commerce Research and Applications* 14 (6), 393–404.

Amichai-Hamburger, Y. and A. Furnham (2007). "The Positive Net." *Computers in Human Behavior* 23 (2), 1033–1045.

Barnes, S. J., H. H. Bauer, M. M. Neumann and F. Huber (2007). "Segmenting cyberspace: a customer typology for the internet." *European Journal of Marketing* 41 (1/2), 71–93.

Bertot, J., E. Estevez and T. Janowski (2016). "Universal and contextualized public services: Digital public service innovation framework." *Government Information Quarterly* 33 (2), 211–222.

Blank, G. and D. Groselj (2014). "Dimensions of Internet Use: Amount, Variety, and Types." *Information, Communication & Society* 17 (4), 417–435.

Boll, F. and P. Brune (2015). "User Interfaces with a Touch of Grey? - Towards a Specific UI Design for People in the Transition Age." *Procedia Computer Science* 63, 511–516.

Borg, K. and L. Smith (2018). "Digital inclusion and online behaviour: five typologies of Australian internet users." *Behaviour & Information Technology* 37 (4), 367–380.

Brandtzæg, P. B. (2010). "Towards a unified Media-User Typology (MUT): A meta-analysis and review of the research literature on media-user typologies." *Computers in Human Behavior* 26 (5), 940–956.

Brandtzæg, P. B. (2012). "Social Networking Sites: Their Users and Social Implications - A Longitudinal Study." *Journal of Computer-Mediated Communication* 17 (4), 467–488.

Brandtzæg, P. B. and J. Heim (2011). "A typology of social networking sites users." *International Journal of Web Based Communities* 7 (1), 28–51.

Brandtzæg, P. B., J. Heim, B. H. Kaare, T. Endestad and L. Torgersen (2005). "Gender Differences and the Digital Divide in Norway – Is there really a Gendered Divide?" In: *Childhoods: Children and Youth in Emerging and Transforming Societies*, Gyldendal/Oslo, p. 427–454.

Brandtzæg, P. B., J. Heim and A. Karahasanović (2011). "Understanding the new digital divide - A typology of Internet users in Europe." *International Journal of Human Computer Studies* 69 (3), 123–138.

Charness, N. and W. R. Boot (2009). "Aging and Information Technology Use." *Current Directions in Psychological Science* 18 (5), 253–258.

Cook, J. E. and C. Doyle (2002). "Working alliance in online therapy as compared to face-to-face therapy: Preliminary results." *Cyberpsychology and Behavior* 5 (2), 95–105.

Cotten, S. R., G. Ford, S. Ford and T. M. Hale (2012). "Internet use and depression among older adults." *Computers in Human Behavior* 28 (2), 496–499.

Coulson, N. S. (2005). "Receiving social support online: An analysis of a computer-mediated support group for individuals living with irritable bowel syndrome." *Cyberpsychology and Behavior* 8 (6), 580–584.

Delello, J. A. and R. R. McWhorter (2017). "Reducing the Digital Divide: Connecting Older Adults to iPad Technology." *Journal of Applied Gerontology* 36 (1), 3–28.

Deutsche Rentenversicherung (2019). *Rentenatlas 2019: Die Deutsche Rentenversicherung in Zahlen, Fakten und Trends.* URL: https://www.deutsche-rentenversicherung.de/SharedDocs/Downloads/DE/Statistiken-und-Berichte/Rentenatlas/2019/rentenatlas_2019_download.pdf?__blob=publicationFile&v=6 (visited on 11/29/2019).

Distel, B. and J. Becker (2017). "All Citizens are the Same, Aren't They? – Developing an E-government User Typology." In: *Electronic Government. EGOV 2017*. Ed. by M. Janssen, K. Axelsson, O. Glassey, B. Klievink, R. Krimmer, I. Lindgren, P. Parycek, H. J. Scholl, & D. Trutnev. Lecture Notes in Computer Science, Vol. 10428, Cham: Springer International Publishing, pp. 336–347.

Federal Employment Agency (2018). *2018 Annual Report by the Federal Employment Agency*. URL:

https://www.arbeitsagentur.de/datei/annual-report-2018_ba045416.pdf (visited on 03/26/2020).
Felgenhauer, A., M. Förster, K. Kaufmann, J. Klier and M. Klier (2019). "Online Peer Groups – a Design-Oriented Approach To Addressing the Unemployment of People With Complex Barriers." In: *Proceedings of the 27th European Conference on Information Systems*. Stockholm/Uppsala.
Feuls, M., C. Fieseler, M. Meckel and A. Suphan (2016). "Being unemployed in the age of social media." *New Media & Society* 18 (6), 944–965.
Fortier, A. and J. Burkell (2018). "Display and control in online social spaces: Towards a typology of users." *New Media & Society* 20 (3), 845–861.
Füller, J., K. Hutter, J. Hautz and K. Matzler (2014). "User roles and contributions in innovation-contest communities." *Journal of Management Information Systems* 31 (1), 273–308.
Garg, R. and R. Telang (2012). "Role of Online Social Networks in Job Search by Unemployed Individuals." In: *Thirty Third International Conference on Information Systems*. Orlando/Florida.
Garg, R. and R. Telang (2017). "To be or not to be linked: Online social networks and job search by unemployed workforce." *Management Science* 64 (8), 3926–3941.
Heim, J., P. B. Brandtzæg, B. H. Kaare, T. Endestad and L. Torgersen (2007). "Children's usage of media technologies and psychosocial factors." *New Media and Society* 9 (3), 425–454.
Helbig, N. C., J. R. Gil-García and E. Ferro (2009). "Understanding the complexity of electronic government: Implications from the digital divide literature." *Government Information Quarterly* 26 (1), 89–97.
Helsper, E. J. (2011). *The Emergence of a Digital Underclass: Digital Policies in the UK and Evidence for Inclusion*. LSE Media Policy Project Series, Media Policy Brief 3. London School of Economics and Political Science Department of Media and Communications.
Helsper, E. J. and B. C. Reisdorf (2017). "The emergence of a "digital underclass" in Great Britain and Sweden: Changing reasons for digital exclusion." *New Media and Society* 19 (8), 1253–1270.
Hill, R., L. Betts and S. Gardner (2015). "Empowerment and enablement through digital technology in the generation of the digital age." *Computers in Human Behavior* 48, 1–23.
Howard, P., L. Rainie and S. Jones (2001). "Days and Nights on the Internet - The Impact of a Diffusing Technology." *American Behavioral Scientist* 45 (3), 383–404.
Hunsaker, A. and E. Hargittai (2018). "A review of Internet use among older adults." *New Media & Society* 20 (10), 3937–3954.
IBM (2019). *TwoStep Cluster Analysis*. URL: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/base/idh_twostep_main.html (visited on 11/29/2019).
Johnson, G. M. and A. Kulpa (2007). "Dimensions of online behavior: Toward a user typology." *Cyberpsychology and Behavior* 10 (6), 773–779.
Kau, A. K., Y. E. Tang and S. Ghose (2003). "Typology of online shoppers." *Journal of Consumer Marketing* 20 (2), 139–156.
Kim, J.-Y. (2018). "A study of social media users' perceptional typologies and relationships to self-identity and personality." *Internet Research* 28 (3), 767-784.
Klier, J., M. Klier, L. Thiel and R. Agarwal (2019). "Power of Mobile Peer Groups: A Design-Oriented Approach to Address Youth Unemployment." *Management Information Systems* 36 (1), 158–193.
König, R., A. Seifert and M. Doh (2018). "Internet use among older Europeans: an analysis based on SHARE data." *Universal Access in the Information Society* 17 (3), 621–633.
Kuhn, P. and H. Mansour (2014). "Is Internet Job Search Still Ineffective?" *Economic Journal* 124 (581), 1213–1233.
Kumar, R. and U. R. Srivastava (2018). Ageing Workforce: Negative Age Stereotypes and their Impact on Older Workers. *International Journal of Research in Social Sciences* 8 (5), 302-312.
Lee, B., Y. Chen and L. Hewitt (2011). "Age differences in constraints encountered by seniors in their use of computers and the internet." *Computers in Human Behavior* 27 (3), 1231–1237.
Lindsay, C. (2005). "Employability, services for unemployed job seekers and the digital divide." *Urban Studies* 42 (2), 325–339.
Lüders, M. and P. B. Brandtzæg (2017). "'My children tell me it's so simple': A mixed-methods approach to understand older non-users' perceptions of Social Networking Sites." *New Media & Society* 19 (2), 181–198.

McQuaid, R. W., C. Lindsay and M. Greig (2004). "'Reconnecting' the Unemployed Information and communication technology and services for jobseekers in rural areas." *Information, Communication & Society* 7 (3), 364–388.

Menold, N. and Bogner, K. (2016). "Design of rating scales in questionnaires." In: *GESIS Survey Guidelines. Mannheim*, Germany: GESIS – Leibniz Institute for the Social Sciences.

Moen, P., E. Kojola and K. Schaefers (2017). "Organizational Change Around an Older Workforce." *Gerontologist* 57 (5), 847–856.

Morris, A., J. Goodman and H. Brading (2007). "Internet use and non-use: Views of older users." *Universal Access in the Information Society* 6 (1), 43–57.

Niehaves, B. and J. Becker (2008). "The Age-Divide in E-Government - Data, Interpretations, Theory Fragments." In: M. Oya, R. Uda, & C. Yasunobu (Eds.), *Towards Sustainable Society on Ubiquitous Networks*. Boston, MA: Springer US, pp. 279–287.

Nolan, A. and A. Barrett (2018). The role of self-employment in Ireland's older workforce. *Journal of the Economics of Ageing* 14.

Nurgalieva, L., J. J. J. Laconich, M. Baez, F. Casati, and M. Marchese (2019). "A systematic literature review of research-derived touchscreen design guidelines for older adults." *IEEE Access* 7, pp. 22035–22058.

Olphert, C., L. Damodaran and A. May (2005). "Towards digital inclusion - engaging older people in the 'digital world.'" In: *Accessible Design in the Digital World Conference 2005*, Dundee.

Ordonez, T. N., M. S. Yassuda and M. Cachioni (2011). "Elderly online: Effects of a digital inclusion program in cognitive performance." *Archives of Gerontology and Geriatrics* 53 (2), 216–219.

Ortega Egea, J. M., M. R. Menéndez and M. V. R. González (2007). "Diffusion and usage patterns of Internet services in the European Union." *Information Research* 12 (2), 1–13.

Podsakoff, P. M., S. B. MacKenzie and N. P. Podsakoff (2012). "Sources of Method Bias in Social Science Research and Recommendations on How to Control It." *Annual Review of Psychology* 63 (1), 539–569.

Pope, M. (2003). "Career Counseling in the Twenty-First Century: Beyond Cultural Encapsulation." *The Career Development Quarterly* 52 (1), 54–60.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Rousseeuw, P. J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics* (20), pp. 53-65.

Selwyn, N., S. Gorard and J. Furlong (2005). "Whose Internet is it?" *European Journal of Communication* 20 (1), 5–26.

Streitberg, B. and J. Röhmel (1986). "Exact distributions for permutation and rank tests: an introduction to some recently published algorithms." *Statistical Software Newsletter* 12 (1), 10–17.

Sun, F., W. Li, L. Jiang and J. Lee (2020). "Depressive symptoms in three Chinese older workforce groups: the interplay of work stress with family and community factors." *International Psychogeriatrics* 32 (2), 217-227.

Suphan, A., M. Feuls and C. Fieseler (2012). "Social media's potential in improving the mental well-being of the unemployed." In: *Exploring the Abyss of Inequalities. WIS 2012.* Ed. by Eriksson-Backa K., Luoma A., Krook E. Communications in Computer and Information Science, Vol. 313, Berlin, Heidelberg: Springer, pp. 10-28.

United Nations (2018). *E-Government Survey 2018: Gearing E-Government to support transformation towards sustainable and resilient societies*. ST/ESA/PAD/SER.E/205. United Nations, Department of Economic and Social Affairs, July 20.

United Nations (2019a). *The Age of Digital Interdependence*. Report of the UN Secretary-General's High-level Panel on Digital Cooperation, June 10.

United Nations (2019b). *World Population Ageing 2019: Highlights*. ST/ESA/SER.A/430. United Nations, Department of Economic and Social Affairs, Population Division, June 17.

Vansteenkiste, S., N. Deschacht and L. Sels (2015). "Why are unemployed aged fifty and over less likely to find a job? A decomposition analysis." *Journal of Vocational Behavior* 90, 55–65.

Vettori, S. (2016). "Digital Divide Among the 50–56 Year-old Workforce in Malaysia – Preparations

for the Elderly Workforce: Case Study of Selangor, Malacca and Negeri Sembilan." In: S. Vettori (Ed.), *Ageing Populations and Changing Labour Markets*. Routledge.

Vošner, H. B., S. Bobek, P. Kokol and M. J. Krečič (2016). "Attitudes of active older Internet users towards online social networking." *Computers in Human Behavior* 55 (A), 230–241.

Wanberg, C. R., R. Kanfer, D. J. Hamann and Z. Zhang (2016). "Age and Reemployment Success After Job Loss: An Integrative Model and Meta-Analysis." *Psychological Bulletin* 142 (4), 400–426.

White, M. (2001). "Receiving social support online: implications for health education." *Health Education Research* 16 (6), 693–707.

Wildevuur, S. E. and L. W. Simonse (2015). "Information and Communication Technology - Enabled Person-Centered Care for the "Big Five" Chronic Conditions: Scoping Review." *Journal of Medical Internet Research* 17 (3), e77.

Witte, J., M. Kiss and R. Lynn (2013). The Internet and social inequalities in the U.S. The Digital Divide: The Internet and Social Inequality in International Perspective. In: Ragnedda, M. and Muschert, G.W. (Eds), *The Digital Divide: The Internet and Social Inequality in International Perspective*, Routledge, Abingdon, pp. 67-84.

Woetzel, J., J. Remes, B. Boland, K. Lv, S. Sinha, G. Strube, J. Means, J. Law, A. Cadena and V. von der Tann (2018). *Smart cities: Digital solutions for a more livable future*. McKinsey Global Institute. URL: https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/smart-cities-digital-solutions-for-a-more-livable-future (visited on 20/03/2020).

Yoo, Y. (2010). "Computing in everyday life: A call for research on experiential computing." *MIS Quarterly* 34 (2), 213–231.

Yu, R. P., N. B. Ellison, R. J. McCammon and K. M. Langa (2016). "Mapping the two levels of digital divide: Internet access and social network site adoption among older adults in the USA." *Information Communication & Society* 19 (10), 1445–1464.

Zawacki-Richter, O., W. Müskens, U. Krause, U. Alturki and A. Aldraiweesh (2015). "Student Media Usage Patterns and Non-Traditional Learning in Higher Education." *International Review of Research in Open and Distributed Learning* 16 (2), 136–170.

Online Appendix: Survey (Translated from German)

Gender: ☐ f ☐ m Age: ☐ years

Which of the following digital media communication applications do you use? *(Multiple choices possible)*

- ☐ E-mail
- ☐ Messaging (e.g., Whatsapp)
- ☐ Video calls (e.g., Skype)
- ☐ Social networks (e.g., Facebook, Instagram, Twitter)
- ☐ Participation in Internet forums
- ☐ Publication of own contents (e.g., blog)

How often do you use digital media in everyday life?

Daily ☐ Several times a week ☐ Once a week ☐ Several times a month ☐ Less often ☐ Never ☐

Which of the following digital media applications do you use for job search? *(Multiple choices possible)*

- ☐ Search engines (e.g., Google, Yahoo)
- ☐ Websites of companies
- ☐ Social networks (e.g., Facebook, Instagram, Twitter)
- ☐ Career networks (e.g., LinkedIn, Xing)
- ☐ Job platforms (e.g., StepStone, Indeed, Monster)
- ☐ I do not use digital media for job search.

In case there are digital media you do not use for job search: What is the main reason for that? *(Choose one single answer)*

- ☐ There is not a sufficiently high level of data security.
- ☐ Else: ☐
- ☐ Ease of use is not sufficiently given.
- ☐ I see no benefit in it.
- ☐ I have not dealt with it yet.

Please assess each of the following statements:

	Fully applies 1	2	3	4	5	Applies not at all 6
Digital media assist me in my job search.	☐	☐	☐	☐	☐	☐
I wish for more digital offers to assist me in my job search.	☐	☐	☐	☐	☐	☐

Imagine, you could wish for some digital offering from the Federal Employment Agency. How important is it to you that this digital offering provides you with the below aspects?

	Not important at all 1	2	3	4	5	Very important 6
I regularly receive relevant information regarding job search.	☐	☐	☐	☐	☐	☐
I can ask all of my questions and get answers.	☐	☐	☐	☐	☐	☐
I can learn from other job-seekers' experiences and share my own experiences.	☐	☐	☐	☐	☐	☐

2.2 Activating Older Unemployed Individuals: A Case Study of Online Job Search Peer Groups

Full Citation:	Sigler, Irina (2021). Activating Older Unemployed Individuals: A Case Study of Online Job Search Peer Groups. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, Virtual Event, 2379 – 2388. https://doi.org/10.24251/HICSS.2021.291
Copyright Note:	Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0), https://creativecommons.org/licenses/by-nc-nd/4.0/ (cf. https://hdl.handle.net/10125/70904)

Activating Older Unemployed Individuals: A Case Study of Online Job Search Peer Groups

Irina Sigler
University of Ulm
irina.sigler@uni-ulm.de

Abstract

Improving re-employment chances for older unemployed individuals is a priority for policymakers around the world. While digital job search interventions have proven beneficial for young and middle-aged individuals, their value to support re-employment at older ages has not been investigated so far. To shed light on the potential of digital interventions to assist older unemployed individuals, we analyze a unique data set from a randomized field study introducing online job search peer groups at the Federal Employment Agency in Germany. Results suggest that online peer groups offer substantial added value compared with traditional job search counseling. Participation in online peer groups significantly increases the number of job applications and job interview invitations. We show that older unemployed individuals are accessible for digital job search assistance and identify online peer groups as a powerful intervention to activate this target group.

1. Introduction

The loss of employment is a traumatic event for anyone, but the consequences are especially devastating for people above the age of 50. Compared to younger people, older individuals suffer stronger financial and psychological losses related to unemployment [1, 2]. This age group has lower chances of returning to work [3, 4], which often leads to long-term unemployment and an early labor market exit [5]. The cumulative effect of these individual disasters is a tremendous societal challenge, as the sustainability of public finances and economic growth is threatened by unemployment in a rapidly growing sector of the worldwide population [5, 6].

To address this issue, research and policy have developed a wide range of active labor market programs to assist older unemployed individuals. Examples include training, job-search counseling, and subsidized employment [5, 7, 8]. Still, these programs are often cost-intensive [5, 7] and their effectiveness is mixed [7, 8, 9]. Since these programs are mostly delivered in-person, the societal value of Information and Communications Technology (ICT) to improve re-employment chances of older unemployed individuals remains untapped [5, 9].

While initial research testifies to the benefits of ICT for finding employment in young and middle-aged individuals [10, 11, 12, 13, 14, 15], little is known about the effectiveness for older unemployed people. To gain insights into the impact of ICT on this target group, we collected and analyzed a unique data set from a randomized field study introducing a digital labor market intervention for unemployed individuals above the age of 50 at the Federal Employment Agency in Germany between February 2019 and March 2020. The intervention consisted of online peer groups, which build on the social support of individuals who share a common issue or need [16] and connect peers in a discussion forum facilitated by digital media [17, 18]. Online peer groups have previously been shown to be effective across a wide range of settings, such as fighting addiction [19], fostering education [20], and even improving re-employment in young job seekers [15].

To the best of our knowledge, our case study is the first to provide quantitative data on the effectiveness of participating in an online peer group for older unemployed individuals. Participating in online peer groups significantly increased job search activities, including job applications and job interviews, as compared to a control group. Furthermore, our results demonstrate that older unemployed individuals actively use digital approaches to find employment. Our contribution to research and practice is twofold. We identify online peer groups as a powerful measure to activate older unemployed individuals. Second, our research provides evidence that older unemployed individuals are responsive to targeted-group-based digital services in the context of job search.

The research presented in this paper is structured as follows. Section 2 illustrates the theoretical background,

URI: https://hdl.handle.net/10125/70904
978-0-9981331-4-0

followed by a description of the research methodology in Section 3. After demonstrating our results in Section 4, we discuss the implications of our findings in Section 5. In Section 6, we reflect on the limitations of our study and provide directions for further research. A summary concludes our paper in Section 7.

2. Theoretical Background

2.1. Problem Context

The consequences of unemployment at a more advanced age are severe on both the individual and the societal level. Job loss has a particularly devastating effect on older individuals who suffer tremendous financial and psychological losses related to unemployment [1, 2]. As older unemployed individuals have comparatively weak chances of finding new employment [3, 4], job loss in this age group is often a prelude to long-term unemployment and early labor market exit [5].

While the definition of older individuals varies between different countries and regions [5], for this research we focus on individuals above the age of 50. This group is shown to experience age-related workplace discrimination [21] and have lower chances of re-entering the labor market [3, 4].

Research and policy have developed a wide range of active labor market programs to help unemployed people re-enter the labor market [5, 7, 8]. To assess the effectiveness of these programs in improving the likelihood of employment, research suggests the following three indicators: increased job search behaviors, advanced job search skills, and high self-efficacy [9, 22, 23].

First, an increase in job search behaviors is associated with higher chances of re-entering the labor market [24]. Such behaviors constitute a set of career-related activities, such as submitting applications or networking [25]. An improvement in job search behavior is often operationalized as job-search intensity, thus, devoting more time and effort to the job search [22, 23]. Job search intensity declines with age [26] and partly explains lower re-employment chances amongst older workers [4]. Changing job search behavior is therefore a promising means for counteracting these trends. Behavioral learning theory suggests social reinforcement and supervision can induce behavioral change and increase job seeking activities [9].

Second, good job search skills are related to higher job search success and are often the immediate outcomes of active labor market programs [9, 22, 23]. More specifically, these include both the knowledge and the ability to conduct a job search effectively, e.g., CV writing skills, interview skills, developing a clear job search strategy [9, 22, 23]. As older individuals are assumed to have a less developed skillset for the job search in general [3, 9] and digital job search tools in particular [3, 5], improving job search skills is particularly important for this target group.

Finally, the theory of planned behavior suggests that positive attitude and perceived behavioral control impact behavioral performance. In the context of job search, this dimension is often operationalized by increased self-efficacy, i.e., the belief of having the capacity to conduct actions to produce a desired outcome [27], which is closely associated with self-assessed job search skills [9]. Prior research shows that high self-efficacy improves chances of re-entering the labor market [27], as it positively influences job search behavior [9]. As self-efficacy declines with age, active labor market programs designed for older job seekers should aim to improve this dimension [28].

2.2. Digital Labor Market Interventions

"Today, fast-evolving technologies have a potential to transform the traditional way of doing things across all functions and domains of government" [29, p. 29]. However, most job search assistance is delivered in-person [5, 9], and the potential of digital technology to serve the unemployed population remains untapped [5, 9]. In the context of re-employment, digital services might be more efficient as they help to reduce costs, enable scaling, and eliminate time and space constraints [14, 30]. Even more importantly, digital services might be more effective in helping unemployed people re-enter the labor market compared to analog job search assistance.

There is initial evidence testifying to the benefits of ICT for finding employment in young and middle-aged individuals. Access to high-speed internet has been shown to improve re-employment rates in Germany [10]. In addition, online job search raised employment chances among young employment seekers in the U.S. by 25% [11]. Furthermore, engaging in social networks increased the number of job leads and interviews in educated, white-collared workers with an average age of 39, who lost their jobs at large organizations across the U.S. during 2010 [12]. Research also suggests that technology-mediated job search interventions can improve labor market prospects. The introduction of a mobile application facilitated the job search and motivated young job seekers in German states, where youth unemployment was above the national average [13]. Also, digital job search assistance increased the number of job interviews in the U.K., for a group with a median age of 36, almost half of them with a university degree [14].

However, little is known about the effectiveness of digital interventions for improving re-employment in older individuals. Contrary to conventional wisdom, initial research suggests that a large share of unemployed individuals above the age of 50 has access to and even desires more digital job search assistance [31]. Further, research suggests that digital media engagement helps older unemployed individuals to better cope with the psychological consequences of job loss. While an investigation of unemployed social media users showed that social media engagement improved well-being across age groups, the effect was especially prominent in older participants [32]. This is consistent with findings in other settings suggesting that digital engagement improves social connectedness in older individuals [33, 34].

Still, the evidence of digital interventions on re-employment indicators in older individuals remains vague. An Australian study analyzing the effects of an online application assistance found that it increased job finding rates among individuals aged 35 - 50; at the same time, the intervention did not improve outcomes for subjects above the age of 50. The authors argued that job seekers above the age of 50 might have a reduced ability to effectively navigate online resources [35]. Another explanation, demonstrated in research on offline job search interventions, might be the fact that older job seekers mostly benefit from targeted interventions [9, 36].

To date, to the best of our knowledge, no study exists that quantifies the impact of digital job search interventions to support re-employment at older ages. Against this backdrop, this study investigates the effect of online peer groups on indicators related to re-employment.

2.3. Online Job Search Peer Groups

Connecting older unemployed individuals via offline peer groups has been shown to increase employment in participants across a diverse range of different occupations (31% manual or unskilled, 33% skilled or clerical, and 36% managerial or professional) [37]; replicated in a following study [38]. During the proliferation of digital media in the 1990s, online peer groups received increasing research attention [39]. Peer groups build on the social support of individuals who share a common issue or need [16]. The online version brings those individuals together in a forum facilitated by digital media [17, 18]. The peers provide mutual support to each other by exchanging advice, information or empathy [40], and thereby foster a “change in the belief, attitude or behavior” [41, p. 138].

Online peer groups provide benefits to diverse age groups, including the elderly [42] and school-age students [15]. Also, they show positive societal impact across a wide range of settings, such as fighting addiction [19], fostering education [20], and even finding employment [15].

Online peer groups increase employment in pupils of middle schools, main schools, and comprehensive schools in Germany and improve attitude, intensity, and maturity of job seeking, thus providing an effective supplement to face-to-face career counseling [15]. In the context of unemployed individuals with complex barriers, such as mental health issues or addiction, effectiveness is less clear. On the one hand, participants in this German study appreciated the design features of the online peer group, e.g., mobile and anonymous communication [43], and showed positive tendencies along several indicators of successful employment, although none of the results were significant [43]. On the other hand, the results also showed a negative tendency for having a clear career strategy, suggesting a potentially distorting effect of participation in an online peer group [43].

Research to date has not investigated the effectiveness of participating in an online peer group in the older unemployed population. To gain insights into the potential of online peer groups to serve unemployed individuals above the age of 50, we analyze a unique data set from a randomized field study introducing online peer groups at the Federal Employment Agency in Germany.

3. Research Methodology

3.1. Case Setting and Subjects

We conducted our randomized field study in cooperation with the Federal Employment Agency in Germany (Bundesagentur für Arbeit). With 156 employment agencies, roughly 600 branch offices, and over 95,000 employees, the Federal Employment Agency is the largest provider of labor market services in Germany. Services include career counseling, employment placing, and financial support.

Between February 2019 and March 2020, a new digital service was introduced to assist labor market prospects for unemployed individuals above the age of 50. The case study was conducted in Baden-Württemberg, the third-largest state in Germany and home to a large automotive and mechanical engineering industry [44]. During the study, the region experienced slowed economic activity [45].

The digital service consisted of online peer groups that supplemented the traditional face-to-face counseling sessions. The peer groups were realized via the instant messaging client Riot.im based on the Matrix

protocol. Each online peer group consisted of about 20 participants and a professional career counselor who acted as a moderator. In addition to facilitating the exchange of text messages, the application also supported document sharing, such as CVs and brochures for career events. All online peer groups had a total duration of three months and were deactivated afterwards; the last group was deactivated end of January 2020.

To measure the effectiveness of the online peer groups, we conducted a pre-test/post-test on both the treatment and the control group, following an established research design [15, 43]. The study consisted of three stages (see Figure 1 for an overview).

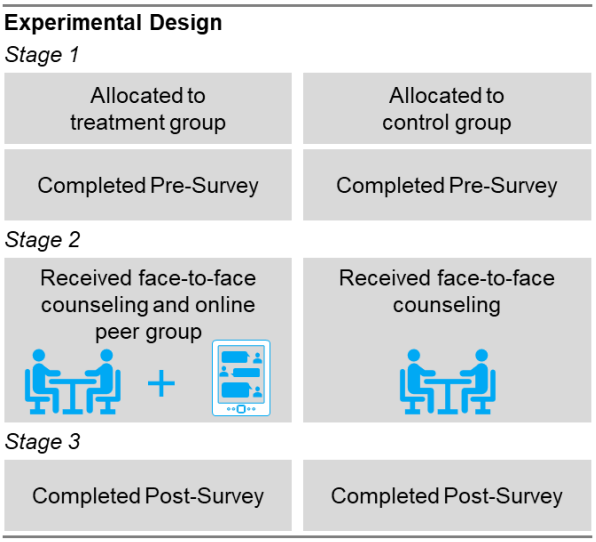


Figure 1. Experimental design

In stage 1, we assessed subjects for eligibility, briefly informed them about the study, randomly allocated them to the treatment and the control group, and requested that they complete a pre-survey. 267 subjects allocated to the treatment and 264 subjects allocated to the control group volunteered to participate and completed the pre-survey.

In stage 2, during three months, all subjects received traditional face-to-face counseling, while subjects in the treatment group also participated in online peer groups. Of the 267 subjects who completed the pre-survey, 205 participated in the online peer groups, and 62 decided to discontinue the project. All 264 subjects allocated to the control group received face-to-face counseling.

In stage 3, we asked all participants to complete a post-survey. 61% of all subjects in the treatment and 47% of all subjects in the control group completed the post-survey. In the analysis, we only included those subjects who completed both surveys. Due to data quality issues, we had to exclude several subjects (5% in the treatment group, 6% in the control group). In total, we collected 119 valid questionnaires for the treatment and 118 valid questionnaires for the control group.

Subjects in the randomized field study were sampled among unemployed individuals above the age of 50 (participants' age-span: 50 - 66) in seven employment agencies, including both rural and urban districts. Most subjects were priorly employed in clerical or professional positions (82%), followed by workers (13%), civil servants (2%), and self-employed individuals (2%). Participation in the experiment was entirely voluntary, with the requirement that all participants have sufficient German language skills to communicate via written messages effectively. Table 1 summarizes the demographics of our sample, based on data gathered from valid pre- and post-surveys.

Table 1. Sample demographics

Variable	N (Percentage)
Gender	
Female	101 (43%)
Male	136 (57%)
Age	
50-55	81 (34%)
55-60	92 (39%)
>60	64 (27%)
Education	
No school leaving certificate	3 (1%)
Lower secondary	36 (15%)
Intermediate secondary	41 (17%)
Upper secondary	17 (7%)
Vocational training	59 (25%)
University degree	79 (33%)
Not specified	2 (1%)
Unemployment duration	
< 3 months	10 (4%)
3-6 months	45 (19%)
6-12 months	74 (31%)
>12 months	108 (46%)

Our analysis is based on two datasets: usage data and survey data. We collected usage data, i.e., all messages, including metadata such as participant identification code and timestamp. In total, 205 unemployed individuals and 11 counselors from 7 employment agencies, replaced by a colleague in case of absence, participated in 11 online peer groups. Second, we conducted pre- and post-surveys with all participants in the treatment and the control group. In the analysis, we only included those subjects who completed both surveys resulting in 119 valid

questionnaires for the treatment group and 118 valid questionnaires for the control group. Participants chose to complete the survey in either a paper-based or online format. The tool "SoSci Survey" was used to collect the online version. All questions were provided in German, and no incentives were offered to respondents.

3.2. Measurement

To ensure reliability and validity of the measurements, we used indicators established by prior research to measure improvement in re-employment chances [9, 22, 23]. To operationalize improvement in job search behavior and self-assessed job search skills, we adopted established constructs from research on the effects of employment interventions in Germany [25]. As digital incompetence impedes job search in older individuals, we added an abbreviated assessment of digital competencies based on established questionnaires [46]. Finally, we used standard questionnaires to measure improvement in self-efficacy [47].

Except for constructs related to the dimension job search behavior, all constructs were measured using a Likert-type scale ranging from 1 ("strongly disagree") to 6 ("agree strongly"). When a construct was comprised of multiple items, we calculated the variable as the mean of the item's score. Table 2 provides an overview of the constructs per dimension, the number of items per construct and the items' scales.

Table 2. Overview on constructs measuring success with respect to employability

Dimension	Construct	Items	Scale
Job search behavior	Number of job applications	1	Free input on number of applications
	Number of invitations to job interviews	1	Free input on number of invitations
Job search skills	Written application skills	4	6-point Likert-type scale
	Career strategy and interview skills	6	
Digital competencies	Information processing	1	6-point Likert-type scale
	Communication	3	
	Safety	2	
	Problem solving	2	
Self-efficacy	Self-efficacy	10	6-point Likert-type scale

The design of our questionnaire follows scientific guidelines to reduce response bias [48, 49]; thus, we selected an even-numbered Likert-type scale to prevent middle option bias [48].

The comprehensibility of the survey items was validated with professionals at the Federal Employment Agency and four Information Systems researchers.

4. Results

4.1. Adoption of Online Peer Groups

First, we determined whether and to what extent participants assigned to the treatment group utilized the digital service by analyzing usage data.

83% of all participants in the treatment group wrote at least one message during the experiment's duration. The participants shared 2,390 messages between each other during the three months of treatment, excluding the messages shared by the professional counselors who acted as moderators (median number of messages in active participants = 8).

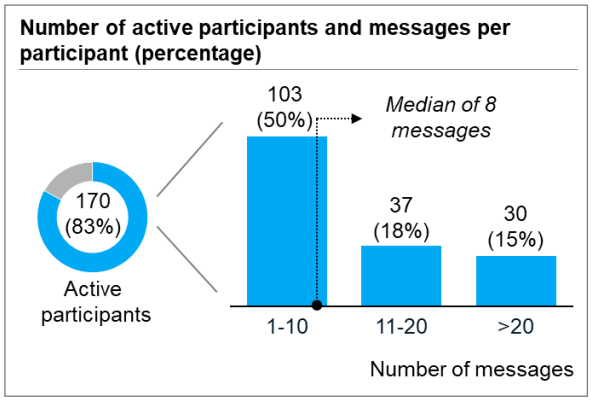


Figure 2. Adoption of online peer groups

4.2. Effect of Online Peer Groups

Comparability of subjects in the treatment and the control group is a prerequisite to attribute any significant difference between the groups to the experimental manipulation, i.e., the online peer groups. Therefore, we verified that the random assignment had indeed produced similar distributions in the treatment and the control group for potentially confounding characteristics such as gender, age, education, German language skills, marital status, children, unemployment duration, and previous occupation [3, 4]. Chi-square analyses of these variables revealed no significant differences between the two groups at the beginning of the experiment.

To evaluate the effect of the online peer-group based intervention, we first derived the differences in each construct per subject in the treatment and the control group. As a next step, we compared the distribution of differences in the control group against the distribution of differences in the treatment group. To investigate whether a significant difference exists in the development of the two groups, we applied Mann-Whitney U tests for independent samples. For this research, a significant difference is defined at a p-value below 0.05.

In what may be the most critical outcome of this study, our results show that job search behavior is positively affected by participating in an online job search peer group.

With respect to the number of job applications, we detect a significant between-group difference according to the Mann-Whitney U test statistic ($p = 0.036$). The median number of job applications of the subjects in the control group does not change between the pre- and post-survey (median control group pre = 10, median control group post = 10). In contrast, the subjects in the treatment group show a striking positive difference, with a median of 10 job applications before the treatment and 14 job applications by the end of the three-month experimental period (see Figure 3).

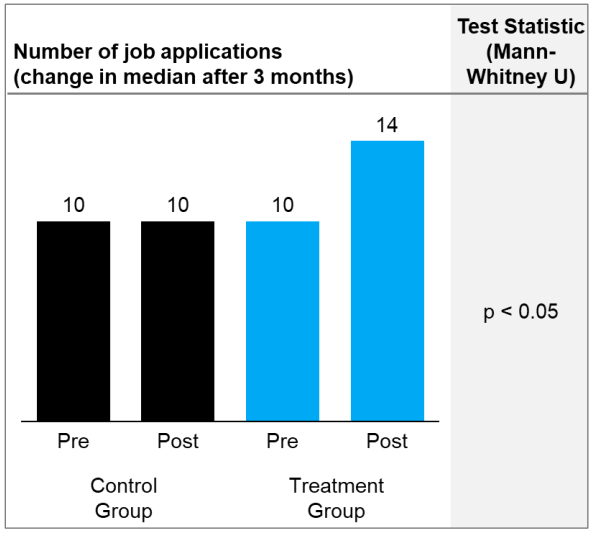


Figure 3. Development in job applications

We also observe that the number of job interview invitations is positively affected by participating in an online peer group (see Figure 4). The subjects in the treatment group show an increased number of invitations to job interviews as compared to control group subjects (Mann-Whitney U test; $p = 0.030$). Similar to the development in the number of job applications, the median number of invitations to job interviews does not change in the control group between the pre- and post-survey (median control group pre = 1, median control group post = 1). Surprisingly, the subjects in the treatment group show a prominent positive difference following the treatment period, with a median of 1 interview invitation before the treatment and 2 invitations afterwards.

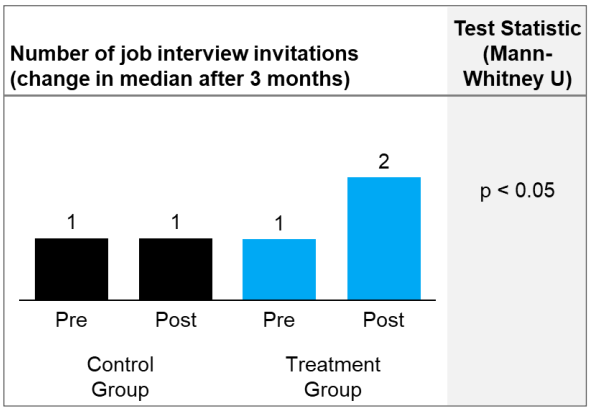


Figure 4. Development in job interview invitations

The treatment group and the control group show no significant difference regarding the development of self-assessed job search skills (see Table 3).

The mean scores for written application skills and career strategy and interview skills do not develop differently following the treatment period between the treatment group and the control group.

Also, subjects in the treatment group and the control group show no difference in the development along the following self-assessed digital skills: information processing, communication, safety, and problem solving (see Table 3).

Finally, there is no significant difference in self-efficacy development between the two groups (see Table 3). The mean self-efficacy scores of the treatment and the control group do not change after the treatment period.

Table 3. Development in job search skills, digital competencies and self-efficacy

Dimension / Construct (6-point Likert-type scale)	**Control Group** (Mean)		**Treatment Group** (Mean)		**Test Statistic** (Mann-Whitney U)
	Pre	***Post***	***Pre***	***Post***	
Job search skills					
Written application skills	4.9	4.9	5.0	5.2	No significant difference
Career strategy and interview skills	4.6	4.6	4.7	4.9	
Digital competencies					
Information processing	4.4	4.5	4.9	5.0	No significant difference
Communication	3.9	4.1	4.0	4.3	
Safety	5.1	5.2	5.3	5.5	
Problem solving	4.7	4.9	5.1	5.3	
Self-efficacy					
Self-efficacy	4.3	4.3	4.5	4.5	No significant difference

5. Discussion

Our research is motivated by the societal challenge of fighting unemployment in older individuals. We investigate whether ICT can meet this challenge and serve to improve the re-employment chances of older job seekers. To this end, we analyzed a unique data set from a randomized field study of online peer groups for unemployed individuals above the age of 50, conducted in cooperation with the Federal Employment Agency in Germany. Our findings have three main implications for theory and practice.

First, our research provides strong evidence that online peer groups help activate older unemployed individuals. We observe a significant improvement in job search behaviors after participation in online peer groups when compared to a control group. The treatment group participants wrote more applications (change in median +4), and were invited to more job interviews (change in median +1) after the treatment. In contrast, the control group shows no improvement in either area. Our research suggests that the change in behaviors observed in offline peer groups [37, 38] can be replicated digitally for the older unemployed population. Also, we confirm that the increase in job search intensity observed in youths [15] can also be demonstrated in older job seekers. Evidence indicates that social support serves as a pathway to increase job search intensity in unemployed individuals above the age of 50 [26, 50]. More specifically, this target group mainly benefits from support provided by unemployed friends, rather than other employed or retired friends or family members [26, 50]. As online peer groups help to connect peers who share a common challenge or need, our research provides preliminary evidence that online peer groups serve as an effective means to build such relationships. The convenience and ubiquitous nature of mobile service delivery might have further facilitated the change in behaviors [51], as research on offline peer groups for older individuals reports a desire for more frequent exchanges with peers [38]. From a practical perspective, activating older job seekers is particularly crucial, as job search intensity declines with age [26] and partly explains lower re-employment chances for older job seekers [4]. In light of this challenge, we identify online peer groups as a powerful measure to increase job search intensity among the older unemployed population. Thus, our research is an initial step towards modernizing labor market services and suggests the adoption of digital peer-group based services in addition to face-to-face counseling.

Second, our results shed light on the limitations of online peer group interventions for older unemployed individuals. We observed no change in written application skills, interview and career strategy skills, digital competencies, and self-efficacy following participation in online peer groups. This contrasts with observations in other settings where online peer groups improved attitude and self-assessed skills [15, 52, 53, 54]. Observations from offline peer groups might shed light on these findings. Older unemployed individuals show improvement in self-efficacy when specific learning assignments complement peer group participation [55]. Other evidence suggests that skill and self-efficacy development require additional components [9, 56], and that peer learning works best as a supplement to other training [57]. Our findings indicate that while peer groups serve the crucial goal of activating older unemployed individuals, additional interventions are needed to improve skills and self-efficacy.

Third, our results demonstrate that older unemployed individuals actively use digital career search assistance. About 83% of all subjects in the treatment group wrote at least one message, and about 33% even wrote over ten messages throughout the duration of the experiment. Thereby, older individuals participate almost as actively as youths in a similar employment-related peer group, which had a 100% participation rate [15]. Further, the participation rate is strikingly high when compared to online peer groups in other settings [19] and large social health networks,

where less than 25% of participants contribute more than one message [58]. The active participation in the online peer groups partly contests findings of an overrepresentation of older unemployed individuals among non-users of digital media [59] and the higher share of older people among "lurkers" in online support groups [60]. Thus, our findings are consistent with research suggesting that older unemployed individuals are accessible for digital assistance in job search [31]. The setting and design of the digital intervention, aimed explicitly at individuals above the age of 50, might serve as a potential reason for the strong participation. Research demonstrates that older job seekers mostly benefit from interventions specifically targeted to them [9, 36]. Furthermore, the intervention design was informed by prior literature. The peer groups were implemented as localized and closed sub-networks [42], aiming at high usability and data protection standards identified as essential features for job seekers above 50 [31]. Our research provides evidence that older unemployed individuals are responsive to targeted-group-based digital services in the context of job search.

6. Limitations and Directions for Future Research

Besides the highlighted research contributions presented in this paper, we acknowledge the limitations of this study that constitute interesting avenues for future research.

First, our data comes from a single case study in one country. Though we build on a rich dataset gathered from several online peer groups at different branch offices of the Federal Employment Agency in Germany, including both rural and urban areas, these findings may not be generalizable to other countries. Thus, we invite future research to investigate online peer groups for older unemployed individuals in other countries to shed light on the impact of cultural and regional differences and to substantiate our findings.

Second, we acknowledge that our initial sample of eligible participants may suffer from self-selection bias. We addressed this bias by the randomized allocation of the participants to the control and the treatment group.

Third, our study provides evidence for introducing an online peer group in addition to face-to-face counseling, as our experiment did not include a no-treatment control group. Thus, we cannot isolate the effects of participating in the online peer group and attending the face-to-face counseling sessions. Still, ethical and legal considerations do not support such a study design in our setting.

Finally, our case study is solely intended as a first step in examining the effect of online peer groups in the older unemployed population. In contrast to prior research, we observe that the increased job seeking activity of older unemployed individuals did not go along with elevated self-efficacy or self-assessed skills. Thus, we suggest future research to further expand our study's findings by investigating the underlying dynamics for the behavior changes in older unemployed individuals. In particular, researchers might investigate the factors influencing the observed activation effect in more detail. Future research might compare cohorts with a diverse educational or professional background, unemployment duration, or chat activity. Also, we invite future research to investigate the inner-workings of online peer groups, e.g., to identify specific moderation types that elevate success.

7. Conclusion

Improving labor market prospects for older people is a priority for policymakers around the world [5], in light of the rapid aging of the worldwide population [5, 6]. To date, research and policy have focused on non-digital job search assistance for this target group [9, 5]. Our research is motivated by the desire to investigate whether ICT can serve the older unemployed population. To this end, we analyzed the effect of online peer groups on several indicators related to re-employment. We conducted the analysis using a unique online peer group data set from a randomized field study with unemployed individuals above the age of 50 at the Federal Employment Agency in Germany.

Our findings suggest that online peer groups can support older unemployed individuals. We observe that online peer groups activate participants and increase the number of job applications and job interviews. Our results further highlight that older unemployed individuals actively participate in online peer groups and thus support the call for a targeted design of digital interventions for this age group.

We hope our findings will encourage further investigation into the power of digital interventions to serve the older unemployed population.

8. References

[1] Klehe, U., J. Koen, and I.E. DePater, "Ending on the scrap heap: The experience of job loss and job search among older workers", in The Oxford handbook of work and aging, Oxford University Press, New York, 2012, pp. 313–340.

[2] A.T. Jebb, M. Morrison, L. Tay, and E. Diener, "Subjective Well-Being Around the World: Trends and Predictors Across the Life Span", Psychological Science, 31(3), 2020, pp. 293–305.

[3] A. Tisch, "The employability of older job-seekers: Evidence from Germany", Journal of the Economics of Ageing, 6(C), 2015, pp. 102–112.
[4] S. Vansteenkiste, N. Deschacht, and L. Sels, "Why are unemployed aged fifty and over less likely to find a job? A decomposition analysis", Journal of Vocational Behavior, 90, 2015, pp. 55–65.
[5] OECD, "Working Better with Age", Paris, 2019, https://www.oecd.org/employment/working-better-with-age-c4d4f66a-en.htm, accessed 6-30-2020.
[6] United Nations, "World Population Ageing 2019", New York, 2019, https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Report.pdf, accessed 6-30-2020.
[7] D. Card, J. Kluve, and A. Weber, "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations", Journal of the European Economic Association, 16(3), 2018, pp. 894-931.
[8] M. Biewen, B. Fitzenberger, A. Osikominu, and M. Waller, "Which Program for Whom? Evidence on the Comparative Effectiveness of Public Sponsored Training Programs in Germany", ZEW Discussion Paper No. 07-042, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim, 2007.
[9] S. Liu, J.L. Huang, and M. Wang, "Effectiveness of Job Search Interventions: A Meta-Analytic Review", Psychological Bulletin, 140(4), 2014, pp. 1–33.
[10] N. Gürtzgen, A. Nolte, L. Pohlan, and G.J. van den Berg, "Do digital information technologies help unemployed job seekers find a job? Evidence from the Internet Expansion in Germany", IZA Discussion Paper No. 11555, Institute for the Study of Labor (IZA), Bonn, 2018.
[11] P. Kuhn and H. Mansour, "Is Internet Job Search Still Ineffective?", The Economic Journal, 124(581), 2014, pp. 1213–1233.
[12] R. Garg and R. Telang, "Role of Online Social Networks in Job Search by Unemployed Individuals", in 33rd International Conference on Information Systems (ICIS), Association for Information Systems, Orlando/Florida, 2012.
[13] A. Felgenhauer, S. Hieronimus, J. Klier, M. Klier, and L. Thiel, "Mobile Job Search Applications–New Pathway to Increase Youths' Job Application Efforts?", in 25th European Conference on Information Systems (ECIS), Association for Information Systems, Guimarães/Portugal, 2017.
[14] M. Belot, P. Kircher, and P. Muller, "Providing Advice to Jobseekers at Low Cost: An Experimental Study on Online Advice", Review of Economic Studies, 86(4), 2019, pp. 1411–1447.
[15] J. Klier, M. Klier, L. Thiel, and R. Agarwal, "Power of Mobile Peer Groups: A Design-Oriented Approach to Address Youth Unemployment", Management Information Systems, 36(1), 2019, pp. 158–193.
[16] Katz, A.H. and E.I. Bender, The strength in us: Self-help groups in the modern world, New Viewpoints, New York, 1976.
[17] T.K. Houston, L.A. Cooper, and D.E. Ford, "Internet support groups for depression: a 1-year prospective cohort study", American Journal of Psychiatry, 159(12), 2002, pp. 2062-2068.
[18] N.S. Coulson, "How do online patient support communities affect the experience of inflammatory bowel disease? An online survey", Journal of the Royal Society of Medicine Short Reports, 4, 2013, pp. 1-8.
[19] J.A. Cunningham, T. van Mierlo, and R. Fournier, "An online support group for problem drinkers: AlcoholHelpCenter.net", Patient Education and Counseling, 70(2), 2008, pp. 193-198.
[20] M. Ambrose, L. Murray, N.E. Handoyo, D. Tunggal, and N. Cooling, "Learning global health: a pilot study of an online collaborative intercultural peer group activity involving medical students in Australia and Indonesia", BMC Medical Education, 17(1), 2017, pp. 1–11.
[21] V.J. Roscigno, S. Mong, R. Byron, and G. Tester, "Age Discrimination, Social Closure and Employment", Social Forces, 86(1), 2007, pp. 313–334.
[22] C.R. Wanberg, L.M. Hough, and Z. Song, "Predictive Validity of a Multidisciplinary Model of Reemployment Success", Journal of Applied Psychology, 87(6), 2002, pp. 1100–1120.
[23] R.W. McQuaid, "Job search success and employability in local labor markets", The Annals of Regional Science, 40(2), 2006, pp. 407–421.
[24] R. Kanfer, T.M. Kantrowitz, and C.R. Wanberg, "Job search and employment: A personality-motivational analysis and meta-analytic review", Journal of Applied Psychology, 86(5), 2001, pp. 837–855.
[25] C. Schmidt, "Wirkungsorientierte Evaluation in der beruflichen Rehabilitation", IQPR Forschungsbericht Nr. 5, IQPR, Cologne, 2007.
[26] J.C. Rife and J.R. Belcher, "Social Support and Job Search Intensity Among Older Unemployed Workers: Implications for Employment Counselors", Journal of Employment Counseling, 30(3), 1993, pp. 98-107.
[27] M. Fugate, A.J. Kinicki, and B.E. Ashforth, "Employability: A psycho-social construct, its dimensions, and applications", Journal of Vocational Behavior, 65, 2004, pp. 14–38.
[28] T.J. Maurer, "Career-relevant learning and development, worker age, and beliefs about self-efficacy for development", Journal of Management, 27(2), 2001, pp. 123–140.
[29] United Nations, "E-Government Survey 2018: Gearing E-Government to support transformation towards sustainable and resilient societies", New York, 2018, https://www.unescap.org/sites/default/files/E-Government%20Survey%202018_FINAL.pdf, accessed 7-14-2020.
[30] R.W. McQuaid, C. Lindsay, and M. Greig, "'Reconnecting' the Unemployed Information and communication technology and services for jobseekers in rural areas", Information, Communication & Society, 7(3), 2004, pp. 364–388.
[31] J. Klier, M. Klier, K. Schäfer-Siebert, and I. Sigler, "#JOBLESS #OLDER #DIGITAL – DIGITAL MEDIA USER TYPES OF THE OLDER UNEMPLOYED", in 28th European Conference on Information Systems (ECIS), Association for Information Systems, An Online AIS Conference, 2020.

[32] Suphan, A., M. Feuls, and C. Fieseler, Social media's potential in improving the mental well-being of the unemployed, in Exploring the Abyss of Inequalities, Springer, Berlin, Heidelberg, 2012, pp. 10-28.
[33] A. Morris, J. Goodman, and H. Brading, "Internet use and non-use: Views of older users", Universal Access in the Information Society, 6(1), 2007, pp. 43–57.
[34] R. Hill, L. Betts, and S. Gardner, "Empowerment and enablement through digital technology in the generation of the digital age", Computers in Human Behavior, 48, 2015, pp. 1–23.
[35] G. Briscese, V. Quinn, and G. Zanella, "Improving Job Search Skills: A Field Experiment on Online Employment Assistance", IZA Discussion Paper No. 13170, Institute for the Study of Labor (IZA), Bonn, 2020.
[36] B. Boockmann and T. Brändle, "Coaching, Counseling, Case-Working: Do They Help the Older Unemployed Out of Benefit Receipt and Back Into the Labor Market?", German Economic Review, 20(4), 2019, pp. e436–e468.
[37] D. Gray, "A job club for older job seekers: An experimental evaluation", Journals of Gerontology, 38(3), 1983, pp. 363–368.
[38] J.C. Rife and J.R. Belcher, "Assisting Unemployed Older Workers to Become Reemployed: An Experimental Evaluation", Research on Social Work Practice, 4(1), 1994, pp. 3–13.
[39] J. Huber, T. Muck, P. Maatz, B. Keck, P. Enders, I. Maatouk, and A. Ihrig, "Face-to-face vs. online peer support groups for prostate cancer: A cross-sectional comparison study", Journal of Cancer Survivorship, 12(1), 2018, pp. 1-9.
[40] N.S. Coulson, "Receiving social support online: An analysis of a computer-mediated support group for individuals living with irritable bowel syndrome", Cyberpsychology and Behavior, 8(6), 2005, pp. 580–584.
[41] W.P. Erchul and B.H. Raven, "Social power in school consultation: A contemporary view of French and Raven's bases of power model", Journal of School Psychology, 35(2), 1997, pp. 137–171.
[42] S. Goswami, F. Köbler, J.M. Leimeister, and H. Krcmar, "Using online social networking to enhance social connectedness and social support for the elderly", in 31st International Conference on Information Systems (ICIS), Association for Information Systems, St. Louis/MO, 2010.
[43] A. Felgenhauer, M. Förster, K. Kaufmann, J. Klier, and M. Klier, "Online Peer Groups – a Design-Oriented Approach To Addressing the Unemployment of People With Complex Barriers", in 27th European Conference on Information Systems (ECIS), Association for Information Systems, Stockholm/Uppsala, 2019.
[44] European Commission, "Regional Innovation Monitor Plus - Baden-Württemberg", Brussels, 2020, https://ec.europa.eu/growth/tools-databases/regional-innovation-monitor/base-profile/baden-w%C3%BCrttemberg, accessed 1-10-2020.
[45] Federal Employment Agency, "Der Arbeitsmarkt 2019 in Baden-Württemberg", Stuttgart, 2020, https://www.arbeitsagentur.de/vor-ort/rd-bw/download/1533733892871.pdf, accessed 1-10-2020.
[46] European Union, "Digitale Kompetenzen - Raster Zur Selbstbeurteilung", Brussels, 2015, https://www.europass-info.de/fileadmin/user_upload/europass-info.de/PDF/Raster_Digitale_Kompetenzen.pdf, accessed 6-30-2020.
[47] W. Mittag and R. Schwarzer, "Interaction Of Employment Status And Self-Efficacy On Alcohol Consumption: A Two-Wave Study On Stressful Life Transitions", Psychology & Health, 8(1), 1993, pp. 77-87.
[48] N. Menold and K. Bogner, "Design of rating scales in questionnaires", GESIS Survey Guidelines, GESIS – Leibniz Institute for the Social Sciences, Mannheim, 2016.
[49] P.M. Podsakoff, S.B. MacKenzie, and N.P. Podsakoff, "Sources of Method Bias in Social Science Research and Recommendations on How to Control It", Annual Review of Psychology, 63 (1), 2012, pp. 539–569.
[50] J.C. Rife, "Older unemployed women and job search activity: The role of social support", Journal of Women & Aging, 7(3), 1995, pp. 55-68.
[51] M. Kleijnen, K. de Ruyter, and M. Wetzels, "An assessment of value creation in mobile service delivery and the moderating role of time consciousness", Journal of Retailing, 83(1), 2007, pp. 33–46.
[52] T.Y. Chung and Y.L. Chen, "Exchanging social support on online teacher groups: Relation to teacher self-efficacy", Telematics and Informatics, 35(5), 2018, pp. 1542–1552.
[53] A. Barak, M. Boniel-Nissim, and J. Suler, "Fostering empowerment in online support groups", Computers in Human Behavior, 24(5), 2008, pp. 1867–1883.
[54] Y.K. Bartlett and N.S. Coulson, "An investigation into the empowerment effects of using online support groups and how this affects health professional/patient communication", Patient Education and Counseling, 83(1), 2011, pp. 113–119.
[55] E.A. Sterrett, "Use of a Job Club to Increase Self-Efficacy: A Case Study of Return to Work", Journal of Employment Counseling, 35(2), 1998, pp. 69-78.
[56] T. Aalbers, M.A. Baars, and M.G. Rikkert, "Characteristics of effective Internet-mediated interventions to change lifestyle in people aged 50 and older: A systematic review", Ageing Research Reviews, 10(4), 2011, pp. 487–497.
[57] K.J. Topping, "Trends in peer learning", Educational psychology, 25(6), 2005, pp. 631-645.
[58] T. Van Mierlo, "The 1% rule in four digital health social networks: An observational study", Journal of Medical Internet Research, 16(2), 2014, pp. 1–9.
[59] M. Feuls, C. Fieseler, M. Meckel, and A. Suphan, "Being unemployed in the age of social media", New Media & Society, 18(6), 2016, pp. 944–965.
[60] C.F. van Uden-Kraan, C.H.C. Drossaert, E. Taal, E.R. Seydel, and M.A.F.J. van de Laar, "Self-reported differences in empowerment between lurkers and posters in online patient support groups", Journal of Medical Internet Research, 10(2), 2008, pp. 1–10.

2.3 Leveraging the Power of Peer Groups for Refugee Integration: A Randomized Field Experiment Comparing Online and Offline Peer Groups

Full Citation:	Förster, Maximilian; Klier, Julia; Klier, Mathias; Schäfer-Siebert, Katharina; & Sigler, Irina (2021). Leveraging the Power of Peer Groups for Refugee Integration. A Randomized Field Experiment Comparing Online and Offline Peer Groups. *Business & Information Systems Engineering*, 1-17. https://doi.org/10.1007/s12599-021-00725-9
Copyright Note:	

Bus Inf Syst Eng
https://doi.org/10.1007/s12599-021-00725-9

RESEARCH PAPER

Leveraging the Power of Peer Groups for Refugee Integration

A Randomized Field Experiment Comparing Online and Offline Peer Groups

Maximilian Förster · Julia Klier · Mathias Klier · Katharina Schäfer-Siebert · Irina Sigler

Received: 6 November 2020 / Accepted: 6 September 2021

Abstract Refugee integration, one long-term solution to the large number of people fleeing their home countries, constitutes a challenge for both refugees and host societies. ICT and especially online peer groups seem promising to support this process. Building on literature demonstrating the societal benefits of peer groups, this paper proposes a novel peer-group-based approach to address refugee integration and introduces both an online and offline realization. A randomized field experiment in cooperation with public (refugee) services and a non-governmental organization makes it possible to expand existing research by quantitatively demonstrating societal benefits of online peer groups and ICT for refugee integration. Further, this paper is the first to assess the effectiveness of online and offline peer groups in one experimental setup comparatively. Results show that peer groups provide substantial value with respect to the integration domains social bridges, social bonds, rights and citizenship as well as safety and stability. While the outcome of the various integration domains differs for online and offline peer groups, participants' adoption rates were higher for online peer groups.

Keywords Online peer group · Refugee integration · Field experiment · Design science

1 Introduction

Humans are born as "ultra-social animals" (Tomasello 2014, p. 187) and started grouping into communities over 50 million years ago (Shultz et al. 2011). Since then, cooperating in groups has been a central strategy for humanity to face challenges. A prominent instrument which builds on this characteristic of human nature are peer groups (Barak et al. 2008). Peer groups differ from other communities (e.g., communities of practice) in such a way that individuals share a need, handicap or desired social/personal change and support each other to overcome their challenging situation or better deal with it (Katz and Bender 1976; Felgenhauer et al. 2019b). Such groups have been proven successful in addressing social problems in various contexts like health (e.g., Cella et al. 1993), career (e.g., Siegel and Donnelly 1978), or racism (e.g., Elligan and Utsey 1999). During the proliferation of the 'social web' in the 1990s, a new variant of peer groups emerged: online peer groups (Huber et al. 2018). Indeed, Information and Communication Technology (ICT) can create enormous societal value among geographically dispersed individuals (United Nations 2019) and contribute towards mitigating the consequences of global crises (Thomas et al. 2020), such as supporting refugee integration (Díaz Andrade and Doolin 2016; 2019). Online peer groups have

Accepted after two revisions by Alexander Maedche.

M. Förster · M. Klier (✉) · K. Schäfer-Siebert · I. Sigler
Institute of Business Analytics, University of Ulm, Helmholtzstraße 22, 89081 Ulm, Germany
e-mail: mathias.klier@uni-ulm.de

M. Förster
e-mail: maximilian.foerster@uni-ulm.de

K. Schäfer-Siebert
e-mail: katharina.schaefer-siebert@uni-ulm.de

I. Sigler
e-mail: irina.sigler@uni-ulm.de

J. Klier
Department of Management Information Systems, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany
e-mail: julia.klier@wiwi.uni-regensburg.de

Published online: 15 November 2021

expanded the applicability of peer groups to various social problems and for instance demonstrated positive effects on individuals in the context of unemployment (e.g., Felgenhauer et al. 2019a) and chronic disease (e.g., Wang et al. 2017). What is more, research postulates that ICT might reinforce support in peer groups; still, research calls for extracting the relative importance of online characteristics in online peer groups (Klier et al. 2019).

To the best of our knowledge, no approach exists to date that exploits the potential of online peer groups to effectively enhance refugee integration, one of today's most pressing issues. The number of refugees, i.e. individuals forcibly displaced due to prosecution, conflict, or general violence, has reached an unprecedented peak of over 25 million worldwide (UNHCR 2020b). Today, integration of this vast number of refugees is a tremendous challenge which confronts both refugees and their host countries. Research indicates that integration of refugees often remains an unsolved issue with refugees risking long-term financial dependency from their host countries, isolation or marginalization as a group, and the hazard of increasing political radicalization in host countries (UNHCR 2013). Even though calls for a "substitute community-type resource" for refugees reach back to the 1980s (Glassman and Skolnik 1984, p. 47), research has rarely dealt with offline peer groups in this context (Badali et al. 2017) and has neglected the societal impact of ICT.

Against this background, we develop a novel peer-group-based approach to enhance refugee integration. We propose a mobile messaging solution (online realization) and a concept for face-to-face meetings (offline realization). Following design science methodology (Hevner et al. 2004), we evaluate the proposed artefact with respect to integration outcomes through a randomized field experiment conducted in cooperation with public (refugee) services and a non-governmental institution. Our contribution to research and practice is threefold. First, we design and implement a novel online peer-group-based approach exploiting the potential of ICT and peer groups in the context of refugee integration. Second, we extend insights into the effects of ICT and online peer groups in the context of refugee integration based on a randomized field experiment, thus answering the call for "more empirically grounded studies" in this context (AbuJarour et al. 2019, p. 15). Third, in a comparative analysis of online and offline peer groups, we quantitatively demonstrate differences in their effectiveness for integration outcomes.

The research presented in this paper is structured as follows: In the next section, we illustrate the problem context and provide an overview of the relevant literature on ICT and peer groups. Afterwards, we propose a novel peer-group-based approach for refugee integration with an online and an offline realization. Then, we demonstrate the practical applicability of our artefact and evaluate its efficacy using a randomized field experiment before we critically discuss implications and limitations of our study and provide directions for further research. Finally, we conclude with a summary of our results.

2 Theoretical Background

2.1 Problem Context

The Geneva Convention defines a refugee as an individual who "owing to well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country" (UNHCR 1951). The consequences of flight and displacement are severe, not least because in the last decade (2010–2019) merely a fraction of the roughly 100 million people forcibly displaced worldwide could find a solution to their situation (UNHCR 2020a). Thus, local integration of refugees plays a highly relevant role as a durable solution of displacement (UNHCR 2020a).

While early scholars equated integration with assimilation into the host society (Park and Burgess 1924), nowadays the UN Refugee Agency describes integration as a concept based on "adaptation" and "welcome" and defines integration along three interlinked dimensions – economic, legal, and social-cultural (UNHCR 2013). Following a modern definition of refugee integration, studies have developed several frameworks and models decomposing the concept of refugee integration into domains or dimensions which show reoccurring key aspects of integration. Harder et al. (2018), for example, differentiate between six dimensions, namely 'psychological', 'economic', 'political', 'social', 'linguistic', and 'navigational'. AbuJarour et al. (2018) differentiate between well-being and a sense of agency and, based on a literature review, identify seven dimensions relevant for agency, i.e. 'social networking', 'employment', 'education and language', 'culture', 'health', 'government and citizenship', and 'housing'. A framework which in great parts corresponds with the framework by AbuJarour et al. (2018) has been proposed by Ager and Strang (2008). This framework is among the most comprehensive models of refugee integration (Hynie et al. 2016) and was developed and verified based on theory and practice, with multiple stakeholders involved (Ager and Strang 2008), and, through its domains, provides "indicators that can be used to evaluate the extent of integration and provide goals for targeting programs" (Hynie et al. 2016, p. 2). Figure 1 shows the ten identified domains related to four overall themes of integration

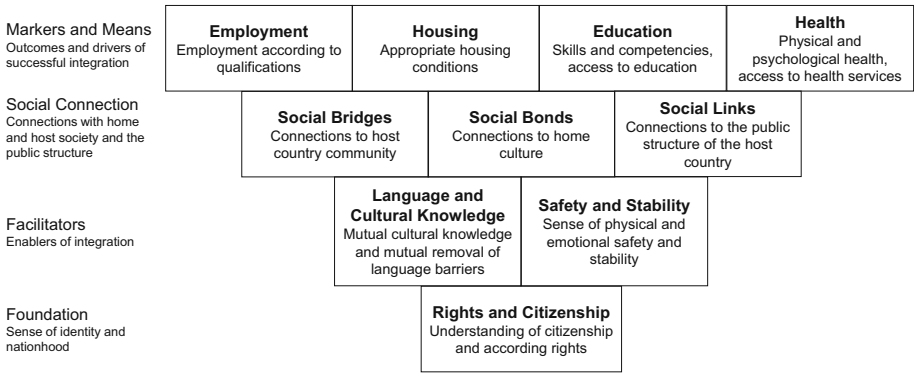


Fig. 1 Integration framework by Ager and Strang (2008)

according to Ager and Strang (2008) which serve as a base for the target and evaluation criteria in our study.

Refugee integration is regarded as a dynamic and two-way process, i.e., involving both refugees and host societies (e.g., Da Lomba 2010; Alencar and Tsagkroni 2019). However, the temporal development of integration varies both across different domains of integration and among individuals according to their individual journeys and experiences (Da Lomba 2010). Further, refugees largely differ in their characteristics and background (AbuJarour et al. 2019). This constitutes an important precondition for the design of refugee services and speaks in favour of highly customizable approaches that can be used for support with respect to a broad range of domains of integration.

2.2 ICT for Refugee Integration

Prior research indicates ICT's potential to help refugees integrating into their host countries (e.g., Siddiquee and Kagan 2006; Bacishoga and Johnston 2013; Díaz Andrade and Doolin 2016; 2019). Mobile phones, for example, have positive effects on social, cultural, and economic participation (Bacishoga and Johnston 2013). Online social networking sites can, for example, serve social connection purposes as well as language and cultural learning purposes (Alencar 2018) and improve women's access to higher education (Dahya and Dryden-Peterson 2017). Digital services constitute a very promising means of supporting refugees. In recent years, many digital services have been introduced for refugees which address different parts of the refugee journey from predeparture over transit, new arrival, and settling, to longer-term integration (Benton and Glennie 2016). So far, there is a focus on short-term issues of refugee integration, i.e., the first time after arrival. Based on the fact that long-term integration is equally important, there is a call of research focusing on these long-term aspects as well (e.g., AbuJarour et al. 2019).

Prior research has designed and evaluated new approaches aiming to support refugee integration in different aspects. The information platform 'Integreat', for instance, offers refugees local information about their municipality by means of different information providers via a mobile application and has been evaluated for optimisation purposes (Schreieck et al. 2017a; 2017b). The mobile application 'Moin' features gamification elements and aims at promoting social events for migrant teenagers as well as providing assistance with contextual language learning in Germany (Ngan et al. 2016). While those examples and many other digital services provide refugees with support from host communities, other digital services provide platforms for refugees to help one another. For example, the health services platform 'New2ukhealth' was designed to provide peer-to-peer support with respect to health issues in the UK (Benton and Glennie 2016), the question and answer (Q&A) site 'Wefugees' provides the opportunity to exchange questions and answers on integration-related topics of all kinds (Schäfer-Siebert and Verhalen 2021), and financial platforms like 'TransferWise' or 'Prosper' allow for peer-to-peer money transfer or lending (Benton and Glennie 2016). One concept which exploits the potential of mutual support among people sharing the same problem or target, are online peer groups (Katz and Bender 1976). So far, research has neglected to investigate online peer groups as an instrument for enhancing refugee integration. However, research on online peer groups in other contexts suggests a high potential of this concept for the purpose of refugee integration.

2.3 Online Peer Groups and Online Peer Group Effects

Peer groups can be defined as networks of people "who have come together for mutual assistance in satisfying a common need, overcoming a handicap or bringing about desired social and/or personal change" (Katz and Bender 1976, p. 278). People in (online) peer groups have been shown to assist each other in various ways which can be grouped into five types of social support: informational support, emotional support, esteem support, network support, and tangible assistance (Cutrona and Suhr 1992).

Due to the proliferation of digital media, online peer groups have received increasing attention in recent years (Huber et al. 2018). In the realm of online communities, online peer groups focus on users that share a challenging situation and pursue to enhance this situation or how to deal with it through mutual support (Katz and Bender 1976; Felgenhauer et al. 2019b; Bedué et al. 2020). In fact, online peer groups have been proven successful in supporting people facing personal and social challenges in different contexts, first and foremost health-related contexts (e.g., Wang et al. 2017), but also in other contexts like parenting (e.g., Niela-Vilén et al. 2014), employment (e.g., Felgenhauer et al. 2019a), and social isolation among elderly (e.g., Goswami et al. 2010). Peer group effects can be defined as a "change in the belief, attitude or behaviour of a person […] which results from the action or presence [of a peer or group of peers]" (Erchul and Raven 1997, p. 138).

Interest in online peer groups has generated a rich literature in diverse contexts revealing a diversity of positive peer group effects. First, peer groups can foster *knowledge gain* by increasing content knowledge through interaction with peers. For instance, parents in online peer groups report to better understand the role of parenting (Niela-Vilén et al. 2014). Second, peer groups can lead to *positive behaviour change* thus altering detrimental practices. For instance, research indicates that a mobile peer-group-based career counselling approach can significantly increase young people's chances of finding employment, while improving their career search intensity (Klier et al. 2019). Third, participants of online peer groups can benefit from an *intensification of social connectedness*, which includes feelings of closeness and belonging to peers (Goswami et al. 2010). For instance, elderly people in online peer groups report to escape social exclusion through increased social participation (Goswami et al. 2010). Beyond this, online peer groups can induce intensification of relationships, especially to professional counsellors. Felgenhauer et al. (2019a), for instance, found that unemployed people with complex employment barriers experienced more target-oriented face-to-face employment counselling if at the same time they participated in an online peer group. A fourth positive peer group effect is an *increase of general well-being*. For instance, online peer groups can induce reductions in depression symptoms for women with postpartum depression (Prevatt et al. 2018). Fifth, peer groups have been found to induce an *increase of self-efficacy*, i.e., the "beliefs in one's capabilities to organise and execute the courses of action required to produce given attainments" (Bandura 1997, p. 3), also referred to as empowerment (Barak et al. 2008) in health-related contexts. For instance, some studies indicate that participation in online peer groups results in improved self-care behaviour of stigmatized chronic diseases (Wang et al. 2017). Apart from those positive effects, some studies also describe unintended side-effects of online peer groups such as the uncritical adoption of potentially harmful information or misinformation (Leist 2013), misuse of personal data (Leist 2013), and harassment under the cloak of anonymity (Cho and Chung 2012).

We expect online peer groups to be an effective means to enhance refugee integration as the five positive peer group effects described above can be directly linked to elements of successful integration (cf. Ager and Strang 2008) and are thus desired outcomes in this context, too. First, refugees need to learn a foreign language and become familiar with a foreign culture (Ager and Strang 2008; OECD/EU 2018). Peer groups might induce this *knowledge gain* (e.g., Niela-Vilén et al. 2014). Second, *positive behaviour change* (e.g., Klier et al. 2019) might contribute to employment, for instance through increased job-search behaviour as could be observed by Klier et al. (2019). Third, *intensification of social connectedness* plays an essential role in integration, as refugees need to keep connections to their home country while building relationships with the people and getting acquainted with the institutions in their host country (Ager and Strang 2008). Online peer groups may foster this connectedness, as they are observed to elevate social participation (e.g., Goswami et al. 2010) and to intensify the relationship to a professional counsellor (Felgenhauer et al. 2019a). Fourth, an *increase of general well-being* (e.g., Prevatt et al. 2018) related to (emotional) safety and stability might be desirable in the context of refugees, as many refugees have experienced violence and persecution. Apart from these parallels between already measured peer group effects in other contexts and domains of successful integration, the peer group effect *increase of self-efficacy* (e.g., Barak et al. 2008) could help refugees along their path of integration. Considering the wide range of challenges for integration, self-reliant coordination between different interventions is indispensable, and a high level of refugees' self-efficacy might thus contribute to a more target-oriented integration (Desiderio 2016).

To sum up, prior research indicates ICT's potential to enhance refugee integration. However, there is a scarcity of research on ICT's potential to assist refugees in integrating into their host countries apart from their first time after arrival. Online peer groups might be promising to enhance refugee integration by means of peer group effects. Despite online peer groups' striking societal value in various contexts, to date no approach exists that exploits the potential of online peer groups to effectively enhance refugee integration. We aim to address this research gap by conducting a design science study.

3 Peer-Group-Based Approach to Enhance Refugee Integration

In the following, we propose a novel peer-group-based approach to enhance refugee integration. Based on literature, we design two variants of this approach: an online and an offline realization. Both realizations are designed for the refugees (in general) as participants in our approach. The artefact primarily aims at improving refugee integration on behalf of the refugees within the two-way process of refugee integration (cf. e.g., Da Lomba 2010; Alencar and Tsagkroni 2019). However, refugees are also peers, thus representing one central component of our artefact. The peer-group-based approach assists refugees by making use of the enormous potential of peer groups demonstrated in literature. Supplementing existing public and non-governmental interventions, online peer groups (realization A) and offline peer groups (realization B) allow a group of refugees who all need to integrate into a host country to exploit the potential of peer support. In conceptualizing our artefact, we made four major design decisions based on prior research (see Fig. 2).

First, we decided that all refugees with a right to stay and basic skills in the host country's language qualify as peers, independently of their age, gender, language, or cultural background. Both conditions, i.e., having a right to stay and possessing basic skills in the host country's language, are linked with a certain duration of stay. This choice of target group is motivated by three main reasons. First, this way, we take up calls for research on phases of integration other than the first time after arrival in the host country (e.g., AbuJarour et al. 2019). Second, this decision ensures that participants share a common challenge (Katz and Bender 1976), i.e., longer-term integration. Consequently, refugees have already gained some experiences in terms of integration challenges, for instance in learning the host country's language, finding employment, navigating bureaucracy, or identifying leisure activities, and thus might provide mutual understanding and better serve as 'experts' for one another (Barak et al. 2008). Finally, the conditions of a right to stay and basic knowledge skills alone still allow for a certain level of heterogeneity within the group which enhances the diversity of knowledge gain and social connectedness within the group (Lyle 2009).

Second, we chose to build small peer groups with each group consisting of at most 20 refugees. This decision is inspired by literature on job clubs suggesting small group sizes (Azrin et al. 1975). Such small group sizes have recently been proven to be effective in the context of job-search among people with complex barriers (Felgenhauer et al. 2019a) and in the context of social support for refugee women (Liamputtong et al. 2016).

Third, we decided that each peer group is moderated by two experts, one professional counsellor from public (refugee) services and one social worker from a non-governmental organization. The moderators' role is to improve the quality and credibility of information, identified as key design criterion in the refugee setting (Schreieck et al. 2017b), to control the spread of misinformation (e.g., Ross et al. 2018), to prevent bullying (Cho and Chung 2012), and to mediate conflicts that might arise due to cultural tensions (Mogire 2016). Moderators do not introduce any additional pedagogical methods to facilitate improvement along any integration domain in order to allow the peers to

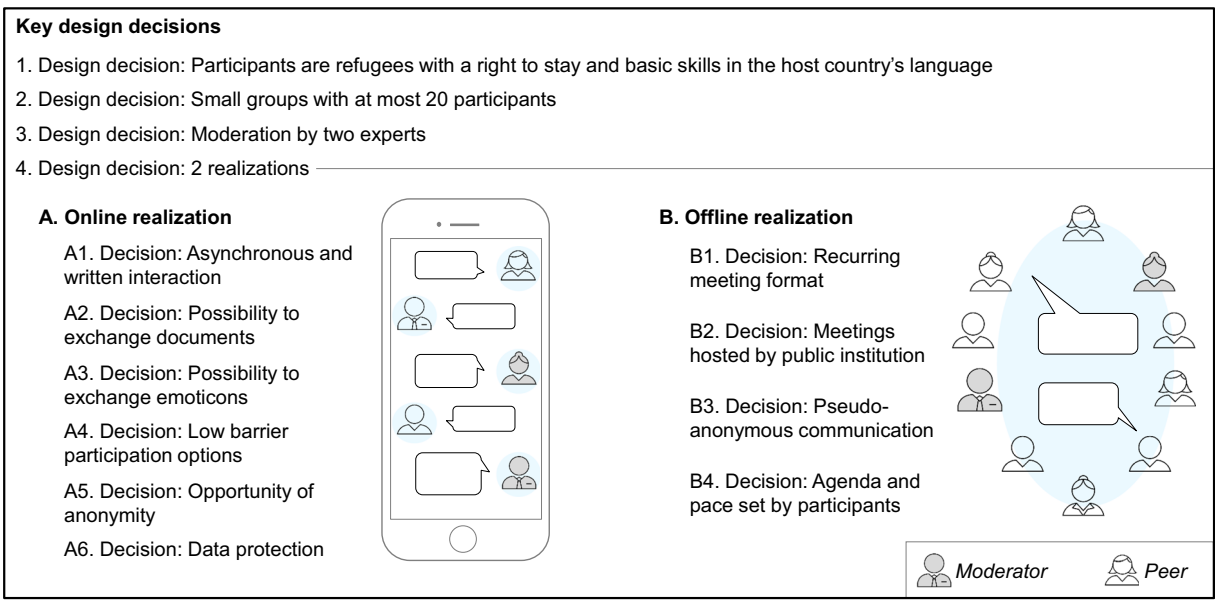


Fig. 2 Online peer groups and offline peer groups to enhance refugee integration

determine the way in which the approach is used. The two types of experts allow for a wider range of competencies: While the professional counsellor from public (refugee) services provides expert knowledge on domains such as *employment*, *education* and *language and cultural knowledge*, and existing public interventions addressing other integration domains like *health* and *housing*, the social worker can provide support on a more diverse range of topics including private housing, culture, daily life, mentoring, and social participation. Together, the moderators make it possible to establish *social links* to existing interventions from public services and civil society (cf. Ager and Strang 2008), thereby satisfying the need for coordination and cooperation among actors in the context of refugee integration (Mason and Buchmann 2016).

Finally, we decided to construct a mobile messenger-based variant (online realization) and a face-to-face variant (offline realization) as we expect both variants to offer advantages in our context. The online realization seems particularly beneficial as literature expects ICT and particularly smartphones to substantially facilitate integration (Díaz Andrade and Doolin 2016) and empower refugees (AbuJarour et al. 2021). Also, research indicates high usage of smartphones among refugees (Betts et al. 2017), suggesting that refugees have similar access to mobile networks as the global population (Vernon et al. 2016). More specifically, mobile connectivity is shown to play a critical role during the migration journey (Dekker et al. 2018; Alencar et al. 2019) and in navigating life in Western host countries (Kaufmann 2018), for example by providing access to education (Drolia et al. 2020). Further, non-copresence, enabled through online communication, renders time and location unimportant and allows for access to support from anywhere and at any time (e.g., Coulson 2013). In our context, a refugee might ask for advice on how to negotiate the contract just before viewing a flat and get immediate support from peers in another city who might have already been in the same situation not long ago. However, online communication also entails disadvantages, as copresence, in contrast, helps people to express attitudes, emotions, and positive appraisal thanks to non-verbal expressions (Kiesler et al. 1985). Consequently, participants might feel closer to each other (Sannomiya and Kawaguchi 1999). This is especially beneficial for our target group as social connection is one factor for successful integration (Ager and Strang 2008).

3.1 Online Realization (A)

In conceptualizing the online realization, we built on literature on online communication and online peer groups to arrive at six (sub-)design decisions as functional requirements that allow to best facilitate integration.

First, we designed our application to build on asynchronous and written interaction, with participants primarily communicating via messages. We chose this interaction mode against the backdrop of refugees communicating in a foreign language and discussing also potentially sensitive topics, as it lowers communication barriers (Braithwaite et al. 1999) and gives participants more time to take up utterances (Andresen 2009). Further, this interaction mode grants participants the flexibility to review older information when needed (Bender et al. 2013).

Second, following the example of Klier et al. (2019), our application allows users to exchange documents beyond simple text messages to foster the exchange of information. In our context, information brochures on integration services, or invitations for job-related events, for example, might be shared.

Third, to facilitate the exchange of emotions and to remedy the absence of non-verbal communication, we decided to allow for exchange of emoticons in our application. We built this decision on literature showing that emoticons facilitate the interpretation of text messages (Derks et al. 2008) and even encourage a caring environment (Klier et al. 2019). Also, such visualizations of text are shown to contribute to a feeling of relaxation and closeness in the context of refugee integration (Kaufmann 2018).

Fourth, to mitigate potential difficulties of communicating in written form in a foreign language, we integrated low-barrier participation options that allow taking part in the conversation without having to formulate a text message, such as conducting a poll.

Fifth, we decided to seize the opportunity of anonymity going along with the feature of non-copresence. Definitions of anonymity largely vary in literature, covering for example namelessness or unidentifiability, and have been shown to be related to both positive and negative types of disinhibition like for example self-disclosure or flaming (Lapidot-Lefler and Barak 2012). For our approach, we decided not to use names but anonymous codes for identification in the groups. This namelessness was established to lower the risk of cultural, religious, or gender-related issues. This way, we further account for the fact that anonymity was identified as a desirable feature by research on online peer groups focusing on sensitive issues, like for example communities for former cancer patients (Bender et al. 2013). Apart from the absence of names, participants were free to share personal information about themselves in the chat conversation. This way, we allowed each participant to control their degree of anonymity as research showed that preferences for anonymity also depend on personal characteristics (e.g., Keipi et al. 2015). We aimed

to counteract potential negative effects of anonymity through moderators being part of each group.

Sixth, we require the application to fulfil additional safeguards securing data protection to lower the risk of misuse of personal data pointed out by prior literature (Leist 2013) and to meet the requirements of data protection in refugee services (Mason et al. 2017).

Apart from these (sub-)design decisions, non-functional requirements ensure the realization of the functionality (cf. Dabbagh and Lee 2015). First, the mobile messaging application needs to be compatible with standard operating systems to allow low-barrier participation. In our case, the messaging application should be compatible with the standard operating systems iOS and Android to potentially reach as many refugees as possible. Second, as a prerequisite to instantiate and manage small online peer groups, the mobile messaging application needs to allow for the creation of closed groups and the invitation of specific users to those groups.

3.2 Offline Realization (B)

In conceptualizing the offline realization, we built on prior literature on offline communication and face-to-face peer groups to arrive at four (sub-)design decisions that allow to best facilitate integration.

First, we decided for a recurring meeting format aiming to establish a positive routine. This decision was guided by literature on job clubs (Azrin et al. 1975), i.e., a context which is also relevant for refugee integration (Ager and Strang 2008), and by literature on peer groups empowering and improving resilience of refugees (Paloma et al. 2020).

Second, we decided for the partnering public (refugee) institution to host all meetings. This way, we aim to foster the linkage between refugees and offered interventions, another important aspect of integration (Ager and Strang 2008), and to lower participation barriers as potential travel expenses can be reimbursed.

Third, we decided to specify pseudo-anonymous communication in that sharing real names was kept optional and that participants could decide themselves for the amount of personal information they share, like in the online setting. This aims to provide an appropriate level of anonymity and privacy facilitating the discussion of sensitive issues (Bender et al. 2013), especially relevant in the context of refugee integration (Paloma et al. 2020).

Fourth, to keep the approach as customizable as possible, the offline realization also serves merely as a space to facilitate mutual support among peers. Thus, the agenda and pace of the meetings are set by the peers themselves, informed by literature on self-help communities (DeCoster and George 2005).

4 Evaluation Strategy

Following design science methodology, we evaluated the utility, quality, and efficacy of our design artefact (Hevner et al. 2004), the peer-group-based approach, and particularly its online and offline realization. We therefore conducted a randomized field experiment and triangled data from three sources to obtain more thorough insights.

4.1 Case Design and Experimental Setting

Conducting a randomized field experiment allowed us to demonstrate the practical applicability of our peer-group-based approach, evaluate its effectiveness and assess online and offline peer groups in the context of refugee integration in a comparative way. The experiment was conducted in cooperation with the German Federal Employment Agency (Bundesagentur für Arbeit) and the German Red Cross at a so-called "Integration Point" in the city of Heidelberg. To respond to a large influx of refugees into Germany since 2015, the Federal Employment Agency instituted "Integration Points" as counselling centres for refugees. The Federal Employment Agency cooperates with municipal authorities and other partners like Employers' Associations to offer a one-stop shop for refugees in these centres. We chose public services counsellors from the "Integration Point" as moderators for our peer-group-based approach as they possess the required expert knowledge required by our design process. We complemented those moderators through a so-called "integration manager" from the German Red Cross according to our third design decision to include a social worker from a non-governmental organization as moderator in our peer groups. These social workers funded by the state usually guide refugees through the large offer of support services and ensure the provision of knowledge on a more diverse range of integration-related topics which is fundamental to the second kind of moderators in our approach.

We sampled subjects for the pilot study among refugees in both rural and urban districts of the "Integration Point". According to our design criteria, we focused on refugees with a right to stay in Germany and with German language skills corresponding to the level B1 of the Common European Framework for Languages to ensure that participants in the peer groups could communicate with each

Table 1 Distribution of the participants' duration of stay

	Min	Max	Median	Mean	Std. Dev
Duration of stay (years)	0.9	9.0	3.4	3.6	0.9

other in German. Participation in the experiment was voluntary. Table 1 demonstrates that participants covered a wide range with respect to their duration of stay. On average, they had been living in Germany for roughly three and a half years at the beginning of the experiment.

The evaluation of our approach is based on a randomized field experiment with two treatment groups using our peer-group-based approach either realized online (online treatment group, T1) or offline (offline treatment group, T2) and a control group (C) receiving traditional counselling. The experiment was conducted in three phases. In the first phase, five voluntarily participating moderators (four professional counsellors from the Integration Point and one counsellor from the German Red Cross), took part in a four-hour workshop to be introduced into their tasks in the peer-group-based approach aiming to establish a common approach to moderation. Acquisition resulted in 196 refugees deciding to participate in the study, with 65 persons in the online treatment group (T1), 63 persons in the offline treatment group (T2), and 68 in the control group (C). Among the participants, there were 59 women and 137 men aged between 18 and 61 years. Most participants (78%) originally came from Syria. Further countries of origin represented in our sample were Iraq, Somalia, Iran, Eritrea, Russia, Turkey, Afghanistan, and China. We asked all 196 participants to complete a pre-survey. Participants in T1 were assisted in installing and introduced to using the messenger immediately after they had decided to participate in the experiment. Participants in T2 received travel expenses when attending the offline meetings. Thus we aimed to ensure that all participants had access to the respective peer group they were offered. In the second phase (three months), participants received support according to their assignment. In the online treatment group (T1), we connected participants of the online peer groups and their respective moderators via the mobile messaging application "Threema Work" as this application meets all (sub-)design decisions and non-functional requirements (cf. Section 3.1) to make it suitable for our artefact (cf. Table 2). Particularly, it allows for the exchange of text messages, documents, pictures, videos, and emoticons and enables low-barrier participation through conducting polls. In contrast to other well-known mobile messaging applications, it is compliant with the EU General Data Protection Regulation (GDPR) and allows for anonymity by usage of randomized identification numbers for participants and deactivated synchronisation between "Threema Work" contacts and private phone books. Compared to the messaging application "Threema", which also meets the design requirements, "Threema Work" particularly qualifies for our experiment, as it additionally allows for a central administration of participants' IDs and the surveillance of their last logins (cf. Section 4.2).

In the offline realization of our approach (T2), the weekly one-hour offline meetings of the participants and their moderators were held at the "Integration Point". The number of groups was chosen such that neither the online peer groups nor the offline meetings exceeded the upper limit of 20 participants determined in our design requirements. The online peer groups and the offline meetings were moderated each by at least one randomly assigned professional counsellor of the "Integration Point" and one social worker from the German Red Cross. The moderators were guided in their moderation tasks by weekly feedback calls and fulfilled the expected role, prevented bullying, added professional knowledge to discussions and shared expert information. Fortunately, there was no need for them to mediate conflicts or to urge participants to be respectful to each other. Online peer groups discussed issues including learning German, finding a job, cultural differences between the home and the host country, leisure activities, and navigating bureaucracy. While these topics were also present in some offline peer group discussions, the latter also included highly intimate topics such as experiences of war and displacement. To help the counsellors in complex situations, we formed a mentoring group using "Threema Work" and instantiated weekly feedback calls with the moderators. In the third phase, we invited all participants again and asked them to complete a post-survey representing the basis for success evaluation. Those who completed the post-survey earned a chance to win regional shopping vouchers worth 15 EUR. We yielded a completion rate of 81% of all 196 participants and counted 54 people in the online treatment group (T1), 53 people in the offline treatment group (T2) and 51 people in the control group (C) who had filled in the pre- and post-survey. Figure 3 summarizes the study design and numbers of participants.

4.2 Data Collection and Measurement

During the experiment, we collected three major datasets: demographic data, usage data, i.e., data on participation in the approach, and survey data.

First, the "Integration Point" provided us with (pseudonymised) demographic data on the participants. This included information on sex, age, country of origin, year of arrival, family status, children, and language level, as these variables have been shown to influence the integration process (Bach et al. 2017). We used this data for robustness purposes.

Second, to capture the adoption of the two realizations of our peer-group-based approach, we gathered data regarding the weekly numbers of participants using the two variants as well as regarding the numbers of participants using the two variants at least once during the three-months

Table 2 Exemplary overview of existing messaging applications and fulfilment of requirements

	Threema Work	Threema	Telegram	ginlo	Wire	Signal	WhatsApp
Asynchronous written interaction mode	✔	✔	✔	✔	✔	✔	✔
Possibility to exchange documents	✔	✔	✔	✔	✔	✔	✔
Possibility to exchange emoticons	✔	✔	✔	✔	✔	✔	✔
Low-barrier participation options	✔	✔	✔	✔	✔	✔	✔
Possibility to remain anonymous	✔	✔	✔	✗	✗	✗	✗
Compliance with GDPR	✔	✔	✗	✔	✔	✔	✗
Availability for iOS and Android	✔	✔	✔	✔	✔	✔	✔
Possibility to create closed groups and invite specific users to those groups	✔	✔	✔	✔	✔	✔	✔

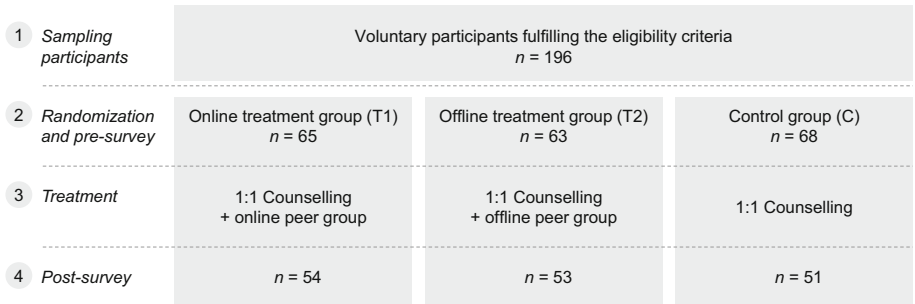


Fig. 3 Study design and numbers of participants

period of the experiment. More precisely, to analyse participants' adoption of the online realization, we collected data on the weekly number of participants using the messenger as well as the number of participants using the messenger at least once during the experiment. This data was gathered by weekly assessing participants' last login times in the messaging application. To analyse participants' adoption of the offline realization, we asked the moderators to track the number of attendants for the offline meetings per week as well as the number of participants attending at least one offline meeting.

Third, we measured individual success with respect to the development of integration domains via pre- and post-surveys. In doing so, we follow common practice in research on the success of Information Systems (IS) (cf. Urbach et al. 2009). The surveys captured items which measure successful integration, based on the integration framework by Ager and Strang (2008). To operationalize the domains of integration by Ager und Strang (2008), we mapped constructs from research on the efficacy of another refugee integration intervention in Germany by Schuller et al. (2011) to the integration domains (cf. Figure 4). A more detailed description of the measurements can be found in the appendix (available online via http://link.springer.com).

4.3 Data Analysis

The purpose of our analysis is twofold. First, we analyse the adoption rates of the two realizations of our approach. Second, we assess the efficacy of the online and offline realization of our peer-group-based approach with respect to the constructs measuring integration success described above.

First, to assess the extent to which people take up the offer of the online and offline peer groups, we calculated the average weekly share of participants using the respective realization (average share of participants using the respective realization at least once) and used Chi-square analyses to test for a significant difference between the online and offline peer group. Second, to determine whether there were significant changes in the online treatment group (T1), the offline treatment group (T2), and the control group (C) during the period of observation with respect to the above described constructs on successful integration, we applied the Wilcoxon signed-rank test to the pre and post values of the constructs of each group. As the only systematic difference between T1, T2, and C is the treatment itself, i.e., the implementation of our peer-group-based approach in the online or offline realization, differences in the developments of the groups should be attributable to our approach. Following similar proceedings

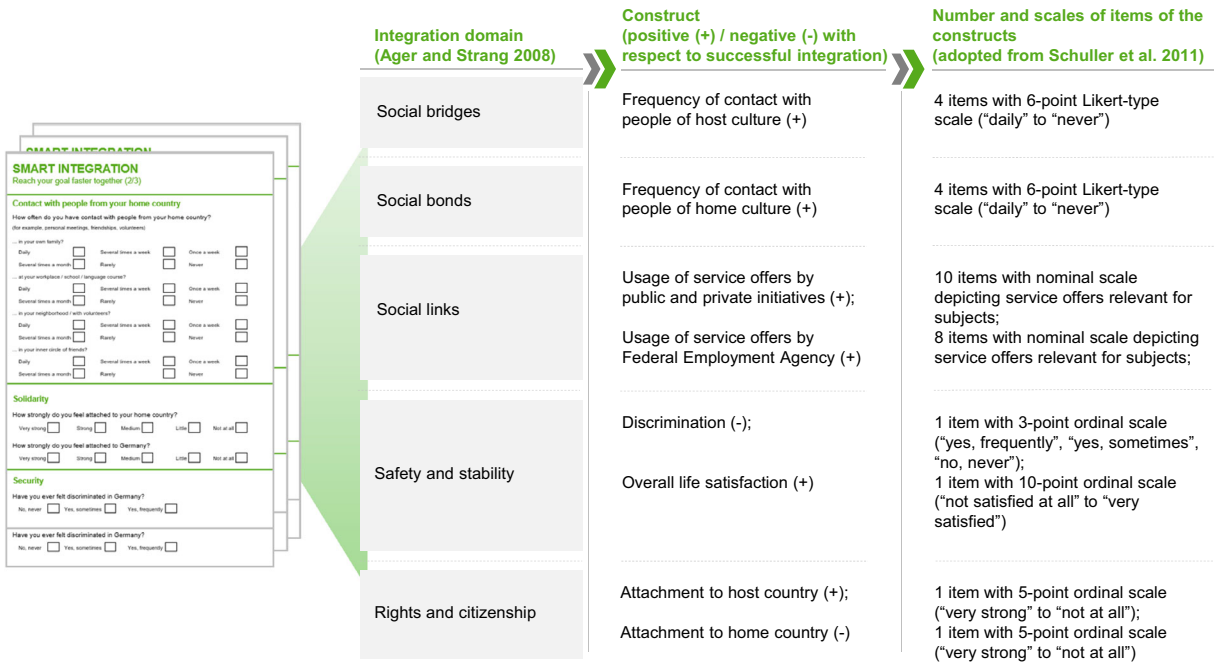


Fig. 4 Overview of analyzed constructs measuring success with respect to integration

in IS literature (e.g., Smith et al. 1998; Im and Hars 2001), we chose the Wilcoxon signed-rank test as a non-parametric alternative to the paired-samples t-test because our data was not normally distributed. For handling zeros, the method by Pratt (1959) was used. *P* values were computed based on the conditional null distribution of the test statistic which was approximated by Monte Carlo resampling. To assure comparability of the three groups, i.e., the online treatment group (T1), the offline treatment group (T2) and the control group (C), and thus to make certain that differences between groups result from the experimental manipulation, we verified the random assignment of participants. To do so, we tested for significant differences in characteristics potentially affecting integration recorded in the demographic data. Chi-square analyses on these variables indicated no significant differences between the three groups at the beginning of the experiment.

5 Results

5.1 Adoption Rates of the Online and Offline Peer Groups

Our first aim was to analyse whether and to what extent participants in the online and offline treatment groups (T1, T2) took up the approach.

As Table 3 shows, the online peer groups were adopted to a higher extent than the offline peer groups.

More precisely, the share of participants in the online treatment group (T1) who visited the online peer groups at least once (70.8%) was higher than the share of participants assigned to test the offline realization (T2) who attended the offline meetings at least once (58.7%). Furthermore, among those participants in the online treatment group (T1), on average 33 participants (50.8%) logged into the messaging application per week (ranging from 11 to 50 participants across weeks, $SD = 10$). In contrast, in the offline treatment group (T2), the share of participants attending an offline meeting was only 7 participants (11.1%) per week on average (ranging from 0 to 17 participants across weeks, $SD = 5$). A Chi-square test of the difference of average share of participants using the two realizations on a weekly basis indicated high significance ($p < 0.001$). While the average number of participants using the approach on a weekly basis reflects regular usage, it does not capture the intensity of usage (e.g., how many messages were sent or read per participant and how intensively participants took part in the discussions of the offline meetings).

5.2 Efficacy of the Online and Offline Peer Groups with Respect to Refugee Integration

Our second aim was to assess the efficacy of the online and offline realization of our peer-group-based approach with respect to refugee integration, decomposed along the integration domains by Ager and Strang (2008). Table 4 gives an overview of the results.

Table 3 Results on the adoption of the online peer groups (T1) and offline peer groups (T2)

	Number of participants	Number (share) of participants using the approach at least once	Average weekly number (share) of participants using the approach
T1	65	46 (70.8%)	33 (50.8%)
T2	63	37 (58.7%)	7 (11.1%)

Table 4 Development of groups (T1, T2, C) with respect to constructs measuring integration success

Constructs and related integration domains (positive (+) / negative (−) with respect to successful integration)		Wilcoxon signed-rank test Z-statistic (*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$)		
		T1	T2	C
Social bridges	(+)	–1.98**increase	–1.85** increase	–1.51* increase
(1) Frequency of contact with people of host culture				
Social bonds	(+)	0.88	–2.28**increase	0.36
(1) Frequency of contact with people of home culture				
Social links	(+)	0.84	0.16	0.51
(1) Usage of service offers by public and private initiatives				
(2) Usage of service offers by Federal Employment Agency	(+)	1.52* increase	0.61	1.88** increase
Safety and stability	(–)	–1.15	–0.69	–1.63* increase
(1) Discrimination				
(2) Overall life satisfaction	(+)	–0.58	1.64* increase	–0,35
Rights and citizenship	(+)	0.65	–2.60*** decrease	–1.46* decrease
(1) Attachment to host country				
(2) Attachment to home country	(–)	−0.49	–0.98	1.76** increase

First, regarding the integration domain social bridges, both the online and the offline treatment groups (T1, T2) significantly improved in the *frequency of contact with people of host culture* ($p < 0.05$). In contrast, the control group (C) only showed an improvement on the 10% significance level. Second, with respect to the domain social bonds, the offline treatment group (T2) showed a significant increase in the *frequency of contact with people of home culture* ($p < 0.05$). In contrast, no significant change in this respect could be detected in the online treatment group (T1) and the control group (C). Third, concerning the domain social links, the control group (C) experienced a significant increase in the *usage of service offers by Federal Employment Agency* on the 5% significance level. While the online treatment group (T1) showed a significant improvement in this respect on the 10% significance level, no such change could be observed in the offline treatment group (T2). Fourth, concerning the domain safety and stability, the control group (C) showed a significant increase in *discrimination* ($p < 0.1$), which could not be observed in the online and offline treatment groups (T1, T2). Further, the offline treatment group (T2) improved significantly with respect to *overall life satisfaction* ($p < 0.1$), whereas the online treatment group (T1) and the control group (C) did not. Finally, regarding the domain rights and citizenship, the control group (C) experienced a significant increase in the *attachment to home country* ($p < 0.05$), while the online and offline treatment groups (T1, T2) did not change significantly. Besides, the control group (C) decreased in the *attachment to host country* on the 10% significance level. Similarly, the offline treatment group (T2) also showed a significant decrease in the *attachment to host country* on the 1% level, whereas the online treatment group (T1) did not show any significant decrease in this respect.

6 Discussion

6.1 Implications for Theory and Practice

Following design science methodology, we developed a novel online peer-group-based approach and an offline realization to enhance refugee integration. We

implemented both the online and the offline realization of the approach in a randomized field experiment to demonstrate the practical applicability of our approach, to evaluate its effectiveness, and to assess the two realizations in a comparative way. The findings contribute to theory and practice in different ways. From a theoretical point of view, they indicate the following three implications.

First, our study provides strong evidence that peer groups provide substantial value to refugee integration in four of five examined domains of integration by Ager and Strang (2008), i.e., *social bridges*, *social bonds*, *rights and citizenship*, and *safety and stability*. Particularly, our study is the first to establish online peer group effects in the context of refugee integration, by means of a randomized field experiment. First, our study shows that peer groups counteract negative developments in refugees' attachment to their home and host country which relates to the peer group effect *positive behaviour change*. While the control group showed both a slightly significant decrease in *attachment to host country* ($p < 0.1$) and an (undesired) significant increase in *attachment to home country* ($p < 0.05$), the online peer groups stayed stable in both of these measures. Studies on online peer groups in other contexts found, for example, an enhancement of participants' attitude towards career choice through online peer groups and eventually their career search intensity (Klier et al. 2019) or positive effects on participants' physical activity mediated by change in intention (Cavallo et al. 2014). While those changes in attitude are closely linked to behaviour, findings in our study concern a general attitude towards a country. Second, we observe an increase of refugees' connectedness to the host country community, i.e., non-peers, which relates to the online peer group effect *intensification of social connectedness* (e.g., Goswami et al. 2010; Felgenhauer et al. 2019a). The construct *frequency of contact with people of host culture* significantly increased in online peer groups ($p < 0.05$) compared to only a slightly significant increase in the control group ($p < 0.1$). While former literature shows online peer groups to go along with improved contact with professionals, for example in the context of unemployment (Felgenhauer et al. 2019a), *intensification of social connectedness* in our study refers to people of the host country in general. This peer group effect is highly relevant in the context of refugee integration, as social connectedness both represents a central dimension in several integration frameworks (cf. e.g., Ager and Strang 2008; Hynie et al. 2016; AbuJarour et al. 2018; Harder et al. 2018) and is explicitly referred to as a target indicator for ICT interventions in this context (e.g., AbuJarour et al. 2019). In demonstrating this peer group effect, our approach stands out from existing integration interventions as they are frequently criticized for isolating refugees (Mason and Buchmann 2016).

Second, our findings highlight that online and offline peer groups when established in the same context and in a comparable way are associated with different peer group effects. While online peer groups in our study provided better outcomes in the integration domain *rights and citizenship*, which relates to the peer group effect *positive behaviour change* (e.g., Klier et al. 2019), they showed weaker outcomes in the integration domains *social bonds* and *safety and stability* which relates to the peer group effects *intensification of social connectedness* (e.g., Goswami et al. 2010) and *increase of general well-being* (e.g., Prevatt et al. 2018), respectively. Both online and offline peer groups showed positive outcomes in the domain *social bridges*. To the best of our knowledge, we are the first to quantitatively demonstrate differences in effectiveness between the two foundational realizations of peer groups: online and offline. We thereby extend understanding of ICT impacts by contributing to the so far unanswered research question of the relative importance of online characteristics in peer groups (Klier et al. 2019). In our study, the following differences were apparent between the two realizations: Only online peer groups stayed stable in the construct *attachment to host country*, whereas offline peer groups showed a highly significant, undesired decrease in that measure ($p < 0.01$). In contrast, there was no significant development in online peer groups with respect to *frequency of contact with people of home culture* and *overall life satisfaction*, whereas offline peer groups significantly increased in both variables as desired ($p < 0.05$; $p < 0.1$). Literature on online characteristics and participants' feedback provides avenues to interpret these differences. Online peer groups are characterized by non-copresence (Coulson 2013). While offline peer groups increased contact with people from their home country, partly by broadening the connection with other refugees in the offline meetings, online peer groups provided support without intensifying contacts amongst each other beyond the participation in the virtual channel. Since online peer group participants only met virtually, they did not strengthen and broaden their network with other refugees, thus, this intervention did not result in increasing their contact to people from their home country. In turn, we conclude that the lower occurrence of a community feeling in the online peer groups allows participants to also feel attached to other people, indicating superior effects with respect to *attachment to host country*. Participants in the offline peer groups reported a different experience with the peer group intervention. They stressed the personal exchange among peers and an atmosphere comparable to a "teahouse", resulting in a feeling of closeness to peers in offline peer groups in line with literature (Sannomiya and

Kawaguchi 1999). Accordingly, prior research suggests that while in online peer groups information plays a more central role, in offline peer groups emotional support and helper therapy are more relevant (Setoyama et al. 2011; Bender et al. 2013). This stronger feeling of connectedness to peers and more central role of helper therapy might explain superior effects of offline peer groups with respect to *frequency of contact with people of home culture* and *overall life satisfaction*.

Through this comparison of online and offline peer groups, we furthermore extend insights into the impact of ICT in the specific context of the study, i.e., refugee integration. Prior studies in this context emphasize the value of ICT with respect to *social bridges* and *social bonds* (e.g., Lloyd and Wilkinson 2017; AbuJarour et al. 2018; Alencar 2018; Kutscher and Kreß 2018). First, while AbuJarour et al. (2018) found that ICT helps resettled refugees to communicate with their friends and family back home and thereby increase their sense of social connectedness, our study suggests that connecting resettled refugees face-to-face is more effective for increasing *social bonds* than connecting them via ICT. Furthermore, existing research proposes that refugees' online communication with people from the host culture is positively correlated with a sense of social connectedness with people from the host culture (AbuJarour et al. 2018). The results of our study expand these findings and suggest that even online communication among refugees themselves can increase *social bridges*. Thus, online peer groups, although 'only' connecting refugees with other refugees, might answer the call for ICT connecting people from the host culture and the home culture (AbuJarour et al. 2019). Finally, prior research found that refugees use ICT to consume and produce cultural content which helps them to maintain a continued connection to their home country (Díaz Andrade and Doolin 2019). In contrast, the online peer groups in our study prove effective for maintaining the *attachment to the host country*: While participation in online peer groups did not increase the *attachment to their home country*, participants in these groups did not experience the decrease of the *attachment to the host country* of the offline peer groups and control group.

Third, our results provide evidence that online peer groups are used to a higher extent than offline peer groups in the context of refugee integration. We find that a significantly higher percentage of participants of the online peer groups (50.8%) used the approach on a regular basis than participants of the offline peer groups (11.1%). While prior research proposes advantages of online peer groups compared to offline peer groups due to time- and location-independent accessibility (Coulson 2013), our study empirically shows that ICT fosters participation in peer groups via a randomized field experiment. In our study, participants reported distance, domestic responsibilities and attending other interventions as main reasons to not make use of the offline peer groups.

Along these theoretical insights, our findings indicate four practical implications to guide decisions in public sector and non-profit organizations.

First, our study demonstrates that peer groups are an effective instrument to enhance refugee integration in four of five dimensions of integration. They particularly help to improve integration by increasing refugees' social connectedness with people from the home and host country and stabilizing their attachment to the home and host country. Against the background that the latest integration summit in Germany (March 2021) reported mixed results with respect to integration interventions for refugee and migrant integration in Germany over the last 15 years, peer groups represent a highly promising approach for refugee integration.

Second, our results show that there is no one-size-fits-all approach to enhance refugee integration, but rather online and offline peer groups are particularly effective in distinct integration domains. Depending on the specific target of integration, the online or offline realization might thus be more advantageous for public sector organizations and non-profit organizations. Being aware of the differences in effectiveness of the two realizations helps organizations to allocate resources more effectively and efficiently.

Third, in the age of digitalisation, the online realization bears advantages for public sector and non-profit organizations. In particular, the online realization of the peer-group-based approach is more promising for implementation on a larger scale. Indeed, our findings regarding the usage of the two realizations suggest that the online realization provides a low-threshold access for participation via smartphone to the peer group as, on average, online peer groups are used more frequently than offline peer groups. At times of crises like Covid-19, online services often remain the only feasible option. The specific insights into online peer group benefits and effects are becoming more relevant as they support stakeholders of public or social services in quickly and reasonably introducing effective digital services whenever necessary.

Finally, organisations that intend to implement a peer-group-based approach to enhance refugee integration should be aware that online peer groups as a digital service demand different working models and competencies than offline peer groups. To illustrate this, moderators of offline peer groups need to host regular in-person meetings (for instance weekly one-hour meetings as in our study), while moderators of online peer groups can flexibly (in time and location) participate in discussions during working hours. This showcases that digitalisation and digital services go

along with different requirements for associated organizations.

6.2 Limitations and Future Research

Aside from the highlighted research contribution presented in this paper, our approach is also subject to limitations which can serve as promising starting points for further research. First, the strengths of our study notwithstanding, our findings are limited regarding the number of participants. Although we could already show significant results for the (separately observed) developments of the two treatment groups and the control group in our study, future research with a larger pool of participants would allow to use more advanced methods to strengthen our results, increase their generality and generate more nuanced insights. For example, methods like differences-in-differences estimators or regression analyses could be used to test for statistical differences between the experimental groups in terms of their development over time. Further, a larger sample would allow for more differentiated insights, e.g., which types of participants extract greater benefit from the online or offline peer groups. Second, the limited observation period of three months did not allow us to analyse long-term effects of our treatments. While we could measure significant developments in domains of integration like *social bonds* and *social bridges* describing refugees' social connectedness, we for instance only found a mitigating effect in *attachment to host country* for online peer groups and could not investigate all integration domains proposed by Ager and Strang (2008). Still, our research provides a promising starting point for future studies investigating long-term effects of online peer groups for refugee integration. Third, despite the valuable opportunity to conduct a field experiment, the generalizability of our findings might be limited by the fact that we conducted our study in one single setting at one "Integration Point". Even though Germany hosts the largest absolute number of refugees among EU countries in mid-2020 (UNHCR 2020b), we invite future research to evaluate our peer-group-based approach in other geographical or cultural settings, as studies on ICT in the context of refugee integration are "a context-specific phenomenon" (Abu-Jarour et al. 2019, p.15). Fourth, in our study, we focused on refugees with basic skills in the home country's language along with a certain duration of stay to maximize the impact of the (online) peer-group-based approach. However, future studies could design variants of this artefact, which allow also new arrivals to participate and benefit from it, and analyse effects on refugee integration for this target group as well. Fifth, even though our artefact primarily focuses on the refugee perspective of the two-way integration process (cf. e.g., Da Lomba 2010; Alencar and Tsagkroni 2019) both in the design and the evaluation of the artefact, professional counsellors from public (refugee) services and social workers from non-governmental organization take part in the approach as moderators and experts. Through participating in the (online) peer groups, those stakeholders potentially learn from the refugees as well. Consequently, there might be positive effects on the host community through the artefact which could be explored in future research. Sixth, our data collection is based on measurement of constructs' initial level and final level to determine the subjects' development in our study. Future research might deepen these insights by observing the continuous development throughout the treatment period, for instance regarding the domain *safety and stability* that may also be subject to more short-term fluctuations. Finally, although we considered two realizations of peer groups for refugees, future studies could conduct another cycle in the iterative design science process (Hevner et al. 2004) and consider further realizations of our artefact, like for example hybrid solutions.

7 Conclusion

Peer groups exploit the social element of human nature and provide an approach that builds on the power of peers to face a shared challenge together, both in face-to-face and online settings. Despite abundant evidence demonstrating online peer groups to be successful in addressing social problems in various contexts, to date no approach exists that exploits the potential of online peer groups in the context of refugee integration, one of today's most pressing issues for both the refugees and their host countries. Further, research calls for assessing the relative importance of ICT in peer groups (Klier et al. 2019).

This study proposed and developed a novel online peer-group-based approach to enhance refugee integration, based on literature on peer groups and ICT effects in peer groups. Besides, we designed an offline realization of the peer-group-based approach. Following design science methodology (Hevner et al. 2004), we evaluated the proposed approach with respect to a well-established framework of integration domains (Ager and Strang 2008) through a randomized field experiment conducted with a unique access at the Federal Employment Agency. Our findings suggest that online peer groups are successful in the integration domains *social bridges, safety and stability*, and *rights and citizenship*. Thus, this research is the first to establish the societal benefits of online peer groups by means of peer group effects in the promising context of refugee integration. Together with promising results for the offline peer groups, we thus provide practitioners with an effective and innovative supplement to existing integration

interventions exploiting the power of peers. Further, our findings indicate that in the context of refugee integration, online and offline peer groups provide better outcomes in different domains of integration: While the online peer groups achieved better effects in the domain *rights and citizenship*, the offline peer group achieved better effects in the domains *social bonds* and *safety and stability*. To the best of our knowledge, we were the first to measure and separately examine peer group effects in online and offline peer groups which have been established in a comparable way in the same context. Thereby, we extend existing understanding of ICT impacts in peer groups. We hope our paper will encourage future research to study the fascinating power of online peer groups.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12599-021-00725-9.

Funding We would like to thank the Péter Horváth-Stiftung for supporting this research.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

AbuJarour S, Krasnova H, Hoffmeier F (2018) ICT as an enabler: understanding the role of online communication in the social inclusion of Syrian refugees in Germany. In: Twenty-Sixth European Conference on Information Systems, Portsmouth

AbuJarour S, Wiesche M, Díaz Andrade A, Fedorowicz J, Krasnova H, Olbrich S, Tan C-W, Urquhart C, Venkatesh V (2019) ICT-enabled refugee integration: a research agenda. Commun Assoc Inf Syst 44:874–891

AbuJarour S, Köster A, Krasnova H, Wiesche M (2021) Technology as a source of power: exploring how ICT use contributes to the social inclusion of refugees in Germany. In: Proceedings of the 54th Hawaii International Conference on System Sciences

Ager A, Strang A (2008) Understanding integration: a conceptual framework. J Refug Stud 21(2):166–191. https://doi.org/10.1093/jrs/fen016

Alencar A (2018) Refugee integration and social media: a local and experiential perspective. Inf Comm Soc 21(11):1588–1603. https://doi.org/10.1080/1369118X.2017.1340500

Alencar A, Kondova K, Ribbens W (2019) The smartphone as a lifeline: an exploration of refugees' use of mobile communication technologies during their flight. Media Cult Soc 41(6):828–844

Alencar A, Tsagkroni V (2019) Prospects of refugee integration in the Netherlands: social capital, information practices and digital media. Media Commun 7(2):184–194

Andresen MA (2009) Asynchronous discussion forums: success factors, outcomes, assessments, and limitations. Educ Technol Soc 12(1):249–257

Azrin NH, Flores T, Kaplan SJ (1975) Job-finding club: a group-assisted program for obtaining employement. Behav Res Ther 13(1):17–27. https://doi.org/10.1016/0005-7967(75)90048-0

Bach S, Brücker H, Haan P, Romiti A, Van Deuverden K, Weber E (2017) Investitionen in die Integration der Flüchtlinge lohnen sich. DIW Wochenbericht 84(3):47–58

Bacishoga KB, Johnston KA (2013) Impact of mobile phones on integration: the case of refugees in South Africa. J Community Inform 9(4):1–12

Badali JJ, Grande S, Mardikian K (2017) From passive recipient to community advocate: reflections on peer-based resettlement programs for Arabic-speaking refugees in Canada. Glob J Community Psychol Pract 8(2):1–31

Bandura A (1997) Self-efficacy: the exercise of control. Freeman, New York

Barak A, Boniel-Nissim M, Suler J (2008) Fostering empowerment in online support groups. Comput Hum Behav 24(5):1867–1883. https://doi.org/10.1016/j.chb.2008.02.004

Beduè P, Förster M, Klier M, Zepf K (2020) Getting to the heart of groups – analyzing social support and sentiment in online peer groups. In: Proceedings of the 41st International Conference on Information Systems, Hyderabad

Bender JL, Katz J, Ferris LE, Jadad AR (2013) What is the role of online support from the perspective of facilitators of face-to-face support groups? A multi-method study of the use of breast cancer online communities. Patient Educ Couns 93(3):472–479. https://doi.org/10.1016/j.pec.2013.07.009

Benton M, Glennie A (2016) Digital humanitarianism: how tech entrepreneurs are supporting refugee integration. Migration Policy Institute, Washington

Betts A, Sterck O, Geervliet R, MacPherson C, Ali A, Memişoğlu F (2017) Talent displaced: the economic lives of Syrian refugees in Europe. Deloitte and the Refugee Studies Centre at the University of Oxford, Oxford

Braithwaite DO, Waldron VR, Finn J (1999) Communication of social support in computer-mediated groups for people with disabilities. Health Commun 11(2):123–151. https://doi.org/10.1207/s15327027hc1102_2

Cavallo DN, Tate DF, Ward DS, DeVellis RF, Thayer LM, Ammerman AS (2014) Social support for physical activity – role of Facebook with and without structured intervention. Transl Behav Med 4(4):346–354. https://doi.org/10.1007/s13142-014-0269-9

Cella DF, Sarafian B, Snider PR, Yellen SB, Winicour P (1993) Evaluation of a community-based cancer support group. Psycho-Oncol 2(2):123–132. https://doi.org/10.1002/pon.2960020205

Cho Y, Chung OB (2012) A mediated moderation model of conformative peer bullying. J Child Fam Stud 21(3):520–529. https://doi.org/10.1007/s10826-011-9538-0

Coulson NS (2013) How do online patient support communities affect the experience of inflammatory bowel disease? An Online Survey JRSM Short Rep 4(8):1–8. https://doi.org/10.1177/2042533313478004

Cutrona CE, Suhr JA (1992) Controllability of stressful events and satisfaction with spouse support behaviors. Commun Res 19(2):154–174. https://doi.org/10.1177/009365092019002002

Dabbagh M, Lee SP (2015) An approach for prioritizing NFRs according to their relationship with FRs. LectNote Softw Eng 3(1):1–5. https://doi.org/10.7763/LNSE.2015.V3.154

Dahya N, Dryden-Peterson S (2017) Tracing pathways to higher education for refugees: the role of virtual support networks and mobile phones for women in refugee camps. Comp Educ 53(2):1–18. https://doi.org/10.1080/03050068.2016.1259877

Da Lomba S (2010) Legal status and refugee integration: a UK perspective. J Refug Stud 23(4):415–436. https://doi.org/10.1093/jrs/feq039

DeCoster VA, George L (2005) An empowerment approach for elders living with diabetes: a pilot study of a community-based self-help group – the Diabetes Club. Educ Gerontol 31(9):699–713. https://doi.org/10.1080/03601270500217787

Dekker R, Engbersen G, Klaver J, Vonk H (2018) Smart refugees: how Syrian asylum migrants use social media information in migration decision-making. Soc Media Soc 4(1):1–11

Derks D, Bos AER, Von Grumbkow J (2008) Emoticons and online message interpretation. Soc Sci Comput Rev 26(3):379–388. https://doi.org/10.1177/0894439307311611

Desiderio MV (2016) Integrating refugees into host country labor markets: challenges and policy options. Migration Policy Institute, Washington DC

Díaz Andrade A, Doolin B (2016) Information and communication technology and the social inclusion of refugees. MIS Q. 40(2):405–416. https://doi.org/10.25300/MISQ/2016/40.2.06

Díaz Andrade A, Doolin B (2019) Temporal enactment of resettled refugees' ICT-mediated information practices. Inf Syst J 29(1):145–174

Drolia M, Sifaki E, Papadakis S, Kalogiannakis M (2020) An overview of mobile learning for refugee students: juxtaposing refugee needs with mobile applications' characteristics. Chall 11(2):31

Elligan D, Utsey S (1999) Utility of an African-centered support group for African American men confronting societal racism and oppression cultural diversity and ethnic minority. Psychol 5(2):156–165. https://doi.org/10.1037/1099-9809.5.2.156

Erchul WP, Raven BH (1997) Social power in school consultation: a contemporary view of French and Raven's bases of power model. J School Psychol 35(2):137–171. https://doi.org/10.1016/S0022-4405(97)00002-2

Felgenhauer A, Förster M, Kaufmann K, Klier J, Klier M (2019a) Online peer groups – a design-oriented approach to addressing the unemployment of people with complex barriers. In: Proceedings of the 27th European Conference on Information Systems, Stockholm

Felgenhauer A, Kaufmann K, Klier J, Klier M (2019b) In the same boat: social support in online peer groups for career counseling. Electron Mark 31(1):197–213. https://doi.org/10.1007/s12525-019-00360-z

Glassman U, Skolnik L (1984) The role of social group work in refugee resettlement. Soc Work Groups 7(1):45–62. https://doi.org/10.1300/J009v07n01_05

Goswami S, Köbler F, Leimeister JM, Krcmar H (2010) Using online social networking to enhance social connectedness and social support for the elderly. In: Proceedings of the 31st International Conference on Information Systems, St. Louis

Harder N, Figueroa L, Gillum RM, Hangartner D, Laitin DD, Hainmueller J (2018) Multidimensional measure of immigrant integration. PNAS 115(45):11483–11488. https://doi.org/10.1073/pnas.1808793115

Hevner AR, March ST, Park J, Ram S (2004) Design Science in Information Systems Research MIS Q 28(1):75–105. https://doi.org/10.2307/25148625

Huber J, Muck T, Maatz P, Keck B, Enders P, Maatouk I, Ihrig A (2018) Face-to-face vs online peer support groups for prostate cancer: a cross-sectional comparison study. J Cancer Surviv 12(1):1–9. https://doi.org/10.1007/s11764-017-0633-0

Hynie M, Korn A, Tao D (2016) Social context and social integration for government assisted refugees in Ontario, Canada. In: Poteet M, Nourpanah S (eds) After the flight: the dynamics of refugee settlement and integration. Cambridge Scholars Publishing, Cambridge

Im I, Hars A (2001) Finding information just for you: knowledge reuse using collaborative filtering systems. In: Proceedings of the 22nd International Conference on Information Systems, New Orleans

Katz AH, Bender EI (1976) The strength in us: self-help groups in the modern world. New Viewpoints, New york

Kaufmann K (2018) Navigating a new life: Syrian refugees and their smartphones in Vienna. Inf Commun Soc 21(6):882–898

Keipi T, Oksanen A, Räsänen P (2015) Who prefers anonymous self-expression online? a survey-based study of Finns aged 15–30 years. Inf Comm Soc 18(6):717–732. https://doi.org/10.1080/1369118X.2014.991342

Kiesler S, Zubrow D, Moses AM, Geller V (1985) Affect in computer-mediated communication: an experiment in synchronous terminal-to-terminal discussion. Hum Comput Interact 1(1):77–104. https://doi.org/10.1207/s15327051hci0101_3

Klier J, Klier M, Thiel L, Agarwal R (2019) Power of mobile peer groups: a design-oriented approach to address youth unemployment. J Manag Inf Syst 36(1):158–193. https://doi.org/10.1080/07421222.2018.1550557

Kutscher N, Kreß LM (2018) The ambivalent potentials of social media use by unaccompanied minor refugees. Soc Media Soc 4(1):1–10. https://doi.org/10.1177/2056305118764438

Lapidot-Lefler N, Barak A (2012) Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. Comput Hum Behav 28:434–443. https://doi.org/10.1016/j.chb.2011.10.014

Leist AK (2013) Social media use of older adults: a mini-review. Gerontol 59(4):378–384. https://doi.org/10.1159/000346818

Liamputtong P, Koh L, Wollersheim D, Walker R (2016) Peer support groups mobile phones and refugee women in Melbourne. Health Promot Int 31(3):715–724. https://doi.org/10.1093/heapro/dav015

Lloyd A, Wilkinson J (2017) Tapping into the information landscape: refugee youth enactment of information literacy in everyday spaces. J Librariansh Inf Sci 51(1):252–259

Lyle D (2009) The effects of peer group heterogeneity on the production of human capital at West Point. Am Econ J Appl Econ 1(4):69–84. https://doi.org/10.1257/app.1.4.69

Mason B, Buchmann D (2016) ICT4Refugees: a report on the emerging landscape of digital responses to the refugee crisis. Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn

Mason B, Schwedersky L, Alfawakheeri A (2017) Digital routes to integration: how civic tech innovations are supporting refugees in Germany. gutorg gemeinnützige Aktiengesellschaft, Berlin

Mogire E (2016) Refugees violent conflict in host states: victims as security threats. Routledge, London

Ngan HY, Lifanova A, Jarke J, Broer J (2016) Refugees welcome: supporting informal language learning and integration with a gamified mobile application. In: Verbert K (ed) Adaptive and adaptable learning. Springer, Cham, pp 521–524. https://doi.org/10.1007/978-3-319-45153-4_54

Niela-Vilén H, Axelin A, Salanterä S, Melender HL (2014) Internet-based peer support for parents: a systematic integrative review. Int J Nurs Stud 51(11):1524–1537. https://doi.org/10.1016/j.ijnurstu.2014.06.009

OECD/EU (2018) Settling in: indicators of immigrant integration. OECD Publishing/European Union

Paloma V, de la Morena I, Sladkova J, López-Torres C (2020) A peer support and peer mentoring approach to enhancing resilience and empowerment among refugees settled in southern Spain. J Community Psychol 48(5):1438–1451

Park RE, Burgess E (1924) Introduction to the science of sociology. University of Chicago Press, Chicago

Poptcheva E, Stuchlik A (2015) Work and social welfare for asylum-seekers and refugees: selected EU member states. PE 572784 European Parliamentary Research Service, Brussels. https://doi.org/10.2861/516230

Pratt JW (1959) Remarks on zeros and ties in the Wilcoxon signed rank procedures. J Am Stat Assoc 54(287):655–667. https://doi.org/10.1080/01621459.1959.10501526

Prevatt BS, Lowder EM, Desmarais SL (2018) Peer-support intervention for postpartum depression: participant satisfaction and program effectiveness. Midwifery 64:38–47. https://doi.org/10.1016/j.midw.2018.05.009

Ross B, Jung AK, Heisel J, Stieglitz S (2018) Fake news on social media: the (in) effectiveness of warning messages. In: Proceedings of the 39th International Conference on Information Systems, San Francisco

Sannomiya M, Kawaguchi A (1999) Cognitive characteristics mediated communication in group discussion an examination from three dimensions. Educ Technol Res. 22:19–25. https://doi.org/10.15077/etr.KJ00003899161

Schäfer-Siebert K, Verhalen N (2021) Thanks for your help! – the value of Q&A websites for refugee integration. In: Proceedings of the 54th Hawaii International Conference on System Sciences

Schreieck M, Wiesche M, Krcmar H (2017a) Governing nonprofit platform ecosystems – an information platform for refugees. Inf Technol Dev 23(3):618–643

Schreieck M, Zitzelsberger J, Siepe S, Wiesche M and Krcmar H (2017b) Supporting refugees in everyday life-intercultural design evaluation of an application for local information. In: Pacific Asia Conference on Information Systems, Langkawi

Schuller K, Lochner S, Rother N (2011) Das Integrationspanel: Ergebnisse einer Längsschnittstudie zur Wirksamkeit und Nachhaltigkeit von Integrationskursen. Forschungsbericht 11 Bundesamt für Migration und Flüchtlinge, Nürnberg

Setoyama Y, Yamazaki Y, Nakayama K (2011) Comparing support to breast cancer patients from online communities and face-to-face support groups. Patient Educ Couns 85(2):e95–e100. https://doi.org/10.1016/j.pec.2010.11.008

Shultz S, Opie C, Atkinson QD (2011) Stepwise evolution of stable sociality in primates. Nat 479:219–222. https://doi.org/10.1038/nature10601

Siddiquee A, Kagan C (2006) The internet, empowerment, and identity: an exploration of participation by refugee women in a community internet project (CIP) in the United Kingdom (UK). J Community Appl Soc Psychol 16(3):189–206. https://doi.org/10.1002/casp.855

Siegel B, Donnelly J (1978) Enriching personal and professional development: the experience of a support group for interns. J Med Educ 53(11):908–914

Smith MA, Mitra S, Narasimhan S (1998) Information systems outsourcing: a study of pre-event firm characteristics. J Manag Inf Syst 15(2):61–93. https://doi.org/10.1080/07421222.1998.11518209

Tomasello M (2014) The ultra-social animal. Eur J Soc Psychol 44(3):187–194. https://doi.org/10.1002/ejsp.2015

Thomas O, Hagen S, Frank U, Recker J, Wessel L, Kammler F, Zarvic N, Timm I (2020) Global crises and the role of BISE. Bus Inf Syst Eng 62(4):385–396. https://doi.org/10.1007/s12599-020-00657-w

UNHCR (1951) Convention relating to the status of refugees. Treaty Series, United Nations, Geneva

UNHCR (2013) A new beginning: refugee integration in Europe. United Nations High Commissioner for Refugees, Geneva

UNHCR (2020a) Global trends: forced displacement in 2019. United Nations High Commissioner for Refugees, Copenhagen

UNHCR (2020b) Mid-year trends 2020. United Nations High Commissioner for Refugees, Copenhagen

United Nations (2019) The age of digital interdependence. Report of the UN Secretary-General's High-level Panel on Digital Cooperation, Brussels

Urbach N, Smolnik S, Riempp G (2009) The state of research on information systems success. Bus Inf Syst Eng 1(4):315–325. https://doi.org/10.1007/s12599-009-0059-y

Vernon A, Deriche K, Eisenhauer S (2016) Connecting refugees: how internet and mobile connectivity can improve refugee well-being and transform humanitarian action. UNHCR, Geneva

Wang X, Parameswaran S, Bagul DM, Kishore R (2017) Does online social support work in stigmatized chronic diseases? A study of the impacts of different facets of informational and emotional support on self-care behavior in an HIV online forum. In: Proceedings of the 38th International Conference on Information Systems, Seoul

Online Appendix

Constructs Measuring Success with Respect to Integration

We measured success with respect to constructs attributable to the integration domains *social bridges*, *social bonds*, *social links*, *safety and stability*, *language and cultural knowledge* as well as *rights and citizenship* which represent foundation, mediators and facilitators of successful integration (Ager and Strang 2008). As language self-assessment in our questionnaire was misunderstood rather as performance test by participants, we decided to discard the according data. We excluded the domains describing markers and means of integration (i.e. achievements and access across the domains *employment*, *education*, *housing*, *health*) (Ager and Strang 2008). Regarding *housing and health*, this is grounded in the fact that these are elements of primary governmental care, with all member states of the European Union being required to provide accommodation and access to healthcare to refugees (Poptcheva and Stuchlik 2015). *Employment* and *education* on the other hand, require a longer observation period, e.g. considering the duration of an application process, and are influenced by other interventions aiming at achievements in and access to *employment* and *education* such as language courses or vocational training provided by the Federal Employment Agency.

We built measurement on the well-established operationalization of integration measures for Germany (Schuller et al. 2011). Comprehensibility of all survey items was validated with professional counsellors of the "Integration Point". Partially, the language of the constructs was simplified and the constructs referring to the domain *social links* were updated to reflect service offers available at the time of the study. Following the recommendation by Schuller et al. (2011), within the scope of this study, those constructs on successful integration consisting of more than one item were aggregated. The average of a constructs' items was realized for the *frequency of contact with people of host culture* (*social bridges*) and for the *frequency of contact with people of home culture* (*social bonds*); the sum of a constructs' items was realized for the *usage of service offers by public and private initiatives* and for the *usage of service offers by Federal Employment Agency* (*social links*).

SMART INTEGRATION

Reach your goal faster together (1/5)

ulm university universität ulm

Questionnaire – *Smart Integration*

Dear participant,

We would like to kindly ask you to complete our questionnaire. Your participation in the survey is anonymous. We will never share your data. The data is only used for science reasons. The declaration of consent applies.

Thank you for taking the time to participate in this survey!

About me

My *nickname* in the study: ☐

Contact with locals

How often do you have contact with Germans/ persons who speak German as their native language?

(for example, personal meetings, friendship, volunteers)

... in your own family?

Daily	☐	Several times a week	☐	Once a week	☐
Several times a month	☐	Rare	☐	Never	☐

... at your working place/ School/ Language course?

Daily	☐	Several times a week	☐	Once a week	☐
Several times a month	☐	Rare	☐	Never	☐

... in your neighborhood/ with volunteers?

Daily	☐	Several times a week	☐	Once a week	☐
Several times a month	☐	Rare	☐	Never	☐

... in your inner circle of friends?

Daily	☐	Several times a week	☐	Once a week	☐
Several times a month	☐	Rare	☐	Never	☐

Solidarity

How strongly do you feel connected to your home country?

Very strong	☐	Strong	☐	Medium	☐	Little	☐	Not at all	☐

How strongly do you feel connected to Germany?

Very strong	☐	Strong	☐	Medium	☐	Little	☐	Not at all	☐

SMART INTEGRATION

Reach your goal faster together (2/5)

ulm university universität ulm

Contact with others from your home country

How often do you have contact with persons from your home country?

(for example, personal meetings, friendship, volunteers)

... in your own family?

Daily	☐	Several times a week	☐	Once a week	☐
Several times a month	☐	Rare	☐	Never	☐

... at your working place/ School/ Language course?

Daily	☐	Several times a week	☐	Once a week	☐
Several times a month	☐	Rare	☐	Never	☐

... in your neighborhood/ with volunteers?

Daily	☐	Several times a week	☐	Once a week	☐
Several times a month	☐	Rare	☐	Never	☐

... in your inner circle of friends?

Daily	☐	Several times a week	☐	Once a week	☐
Several times a month	☐	Rare	☐	Never	☐

Security

Have you ever felt discriminated in Germany?

Yes, frequently ☐ Yes, sometimes ☐ No, never ☐

Counseling Services

Which of the following counseling services have you already used? (You can choose multiple answers)

Career counseling	☐	Integration manager	☐
Youth migration service / Internationaler Bund (IB)	☐	Job information center (BIZ)	☐
Migration advice German Red Cross / Caritas	☐	Asylum groups-voluntary groups	☐
Counseling for recognition of foreign professional qualification IKUBIZ	☐	Counseling offers of the Chamber of Industry and Commerce (IHK)	☐
Counseling offers of the Handelskammer (HWK) Mannheim	☐	Other:	☐

Which of the following counseling services of the “Jobcenter“ have you already used? (You can choose multiple answers)

Entry-level vocational qualification (EQ / long-time internship)	☐	Language and competence check BBQ	☐
KomBer	☐	Mobile Integration Support (TERTIA)	☐
Job- and skills	☐	MySkills	☐
KompAS	☐	Other:	☐

SMART INTEGRATION

Reach your goal faster together (3/5)

ulm university universität ulm

Language Test

We want to learn even more about your German language skills.

Please assess your knowledge of German with the help of the following lists.

There are three lists with different difficulty levels.

Please mark with each task, whether you already can or can't do that.

List 1

I can...

	I can do that	I can't do that
Reading		
... read and understand **timetables** (e.g., bus / train)	☐	☐
... read and understand **street signs** and simple public notices	☐	☐
... read and understand **opening hours** (e.g., stores)	☐	☐
... read and understand a written **appointment**	☐	☐
... read and understand simple **messages**	☐	☐
... read and understand simple, written **directions**	☐	☐
Take part in a conversation		
... **greet** others and **introduce myself**	☐	☐
... ask how to say something in **German**	☐	☐
... have a simple **conversation,** when it comes to a topic that interests me	☐	☐
... ask basic **questions** (e.g., in stores)	☐	☐
... ask for easy **directions** and give them	☐	☐
... lead a simple conversation on the **phone**	☐	☐
Writing		
... fill in a **form** with information about myself	☐	☐
... write down a short **message** (e.g., a sticky note)	☐	☐
... write a simple **postcard**	☐	☐
... write a **greeting card**	☐	☐
... describe **myself** in simple sentences	☐	☐

SMART INTEGRATION

Reach your goal faster together (4/5)

ulm university universität ulm

List 2

I can...

	I can do that	I can't do that
Reading		
... read **operating instructions**	☐	☐
... understand most **advertisements** (e.g., newspapers, magazines)	☐	☐
... look for basic **information** (e.g., on the internet)	☐	☐
... understand **regulations** (e.g., warning signs)	☐	☐
... understand **questions on forms**	☐	☐
... understand short **letters** or information sheets	☐	☐
Take part in a conversation		
... lead a short **conversation** to a familiar topic	☐	☐
... get accurate **information** about something that interests me	☐	☐
... when **shopping** ask for a specific size, color, etc.	☐	☐
... explain a medical issue to my **physician**	☐	☐
... politely express my **approval** or **disapproval**	☐	☐
... express and understand **invitations, apologies and requests**	☐	☐
Writing		
... enter the required information in a **public authority form** or a **questionnaire**	☐	☐
... write down the essential points in a **conversation / telephone conversation**	☐	☐
... write a short and straightforward **report about an event**	☐	☐
... write about me and my **everyday life** (family, school, hobbies)	☐	☐
... write a short **letter** asking for information	☐	☐

SMART INTEGRATION

Reach your goal faster together (5/5)

ulm university universität ulm

List 3

I can...

	I can do that	I can't do that

Reading

... quickly scan a **newspaper** report and understand the gist	☐	☐
... read and understand **public announcements** (e.g., leaflets, community news, instruction manuals)	☐	☐
... understand the **plot in narratives** (e.g., book)	☐	☐
... read and understand **work and study-related texts**	☐	☐
... read **information** and explain it to another person	☐	☐
... understand **letters** from friends	☐	☐

Take part in a conversation

... actively participate in a **discussion** on familiar topics	☐	☐
... speak fluently about **myself**, my **family**, my **interests**, my **job**	☐	☐
... express my **ideas and views** exactly	☐	☐
... ask questions about **topics** that are not **commonplace**	☐	☐
... discuss topics that are reported in **newspapers** and on **television**	☐	☐
... expressing **emotions** (e.g., joy, sadness, interest) in conversation	☐	☐

Writing

... gather details about a **resume**	☐	☐
... write a **short text** on an interesting topic	☐	☐
... respond to an **advertisement** in writing and ask questions	☐	☐
... write or answer a formal or **official letter**	☐	☐
... write a **private letter** to a friend	☐	☐

General satisfaction

How satisfied are you with your life right now?

0	1	2	3	4	5	6	7	8	9	10
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Unsatisfied					Medium					Satisfied

Thank you for your time and all the best for the future!

3. Research on Societal Challenges Posed by Digital Technologies

3.1 Evaluating Explainable Artificial Intelligence – What Users Really Appreciate

Full Citation:	Förster, Maximilian; Klier, Mathias; Kluge, Kilian; & Sigler, Irina (2020). Evaluating Explainable Artificial Intelligence – What Users Really Appreciate. In *Proceedings of the 28th European Conference on Information Systems*, Virtual Conference, 1-18. https://aisel.aisnet.org/ecis2020_rp/195/
Copyright Note:	Reprinted according to author's rights.

6-15-2020

Evaluating Explainable Artifical Intelligence – What Users Really Appreciate

Maximilian Förster
University of Ulm, maximilian.foerster@uni-ulm.de

Mathias Klier
University of Ulm, mathias.klier@uni-ulm.de

Kilian Kluge
University of Ulm, kilian.kluge@uni-ulm.de

Irina Sigler
University of Ulm, irina.hardt@uni-ulm.de

EVALUATING EXPLAINABLE ARTIFICIAL INTELLIGENCE – WHAT USERS REALLY APPRECIATE

Research paper

Förster, Maximilian, University of Ulm, Ulm, Germany, maximilian.foerster@uni-ulm.de

Klier, Mathias, University of Ulm, Ulm, Germany, mathias.klier@uni-ulm.de

Kluge, Kilian, University of Ulm, Ulm, Germany, kilian.kluge@uni-ulm.de

Sigler, Irina, University of Ulm, Ulm, Germany, irina.sigler@uni-ulm.de

Abstract

Explainable Artificial Intelligence (XAI) is an aspiring research field addressing the problem that users of AI do not trust AI systems that act as black boxes. However, XAI research to date is often criticized for not putting the user in the center of attention. We develop a generic and transferable human-based study to evaluate explanations generated by XAI methods from the users' perspective. The design of the study is informed by insights from social sciences into how humans construct explanations. We conduct the study with 164 participants evaluating contrastive explanations generated by representative XAI methods. Our findings reveal characteristics of explanations users appreciate in the context of XAI. We find concreteness, coherence, and relevance to be decisive. These findings provide guidance for the design and development of XAI methods.

Keywords: Explainable Artificial Intelligence, User Study, Contrastive Explanations

1 Introduction

Artificial Intelligence (AI) is increasingly employed for a variety of tasks, first and foremost decision-making and decision support (e.g., Lee, 2018; Lorica and Nathan, 2018; Whittaker et al., 2018). Indeed, AI has the potential to annually create trillions of dollars of value in the global economy (Chui et al., 2018). The arguably most significant impediment for AI deployment is the fact that many AI systems are opaque – or "black boxes" – which means that the reasons for their decisions remain obscure to the user (e.g., Wachter et al., 2018; Guidotti et al., 2019). Opacity fosters users' distrust (Ribeiro et al., 2016) and "if the users do not trust a model or a prediction, they will not use it" (Ribeiro et al., 2016, p. 1135). One prominent example is IBM's "Watson for Oncology", an AI system to assist in cancer treatment whose deployment to hospitals failed as oncologists did not trust it (Bloomberg, 2018). In light of this challenge, the research field of Explainable Artificial Intelligence (XAI) aims to provide methods to automatically generate explanations for the output of AI systems (Gunning, 2017; Kuang, 2017; Abdul et al., 2018; Lipton, 2018). In the context of XAI, an explanation is a human-understandable line of reasoning for why a given input is mapped to an output (Chakraborty et al., 2017; Abdul et al., 2018). Whereas considerable progress has been made in developing new XAI methods, research to date is criticized for not putting the users in the center of attention and not investigating which kind of generated explanations they really appreciate (Miller et al., 2017; Kirsch, 2018; Mittelstadt et al., 2019).

In this paper, we aim to fill this gap by developing and conducting a novel human-based study, building on existing XAI approaches and insights from social sciences. Results suggest that findings from social sciences regarding appreciated characteristics of explanations do not transfer directly to the XAI context. We find that users of XAI appreciate explanations due to the decisive characteristics *concreteness*, *coherence*, and *relevance* in the sense of providing causes of interest, whereas social sciences argue for *shortness* and *generality* instead of *concreteness*. We further identify that the characteristic *length* is

often co-occurring with *concreteness* and that *shortness* is often co-occurring with *relevance*, *generality*, and *consistency*. Finally, we identify potential improvements for the major representative XAI methods employed in the study. Our contribution to theory and practice is threefold. First, we derive and validate a set of decisive characteristics of explanations in the context of XAI. Second, we offer guidance for the user-centric design and development of XAI methods. Third, our work provides a generic study setup for the evaluation of XAI methods from the users' perspective. The remainder of our paper is structured as follows: Section 2 illustrates the problem context. In Section 3, we review relevant literature on explanations in AI and conclude with the research gap. Section 4 covers insights from social sciences on which our study is based and introduces the study design. After presenting our findings in Section 5, we discuss theoretical and practical implications, reflect on limitations of our work, and provide directions for further research in Section 6. A brief summary concludes our paper.

2 Problem Context

AI systems need explanations, as opacity fosters users' distrust (Ribeiro et al., 2016) and distrust reduces people's willingness to accept recommendations, consequently limiting the potential of high-performant but opaque AI systems (Freedy et al., 2007; Herse et al., 2018). Explanations do not merely contribute to gain acceptance (Ye and Johnson, 1995; Benbasat and Wang, 2005; Herse et al., 2018), but also prevent users from blindly following an AI system (Rader and Gray, 2015; Chakraborty et al., 2017). Indeed, explanations should not be exploited to persuade users to follow algorithmic decisions (Gilpin et al., 2018), but rather enable users to "appropriately trust" (Gunning, 2017) an AI system's recommendations (Angelino et al., 2018; Gilpin et al., 2018). This is a critical component towards achieving human agency and thus building a human-centric and "trustworthy" AI (European Commission, 2019). In light of these challenges, the research field of XAI aims to provide methods for automatically generating explanations – human-understandable lines of reasoning – for the output of an AI system (Gunning, 2017; Kuang, 2017; Abdul et al., 2018; Lipton, 2018). To this end, researchers refer to how humans construct explanations themselves. Humans primarily provide contrastive explanations: People focus on decisive aspects, explaining the contrast to an alternative outcome (Lipton, 1990). For instance, a bank advisor would not explain the rejection of a new credit line to a customer by referring to all customer attributes, but rather state that the customer's income or savings would need to be a certain amount higher to obtain approval. In line with these findings, literature suggests contrastive explanations to be a promising concept for XAI (Lim et al., 2009; Dhurandhar et al., 2018; Hoffman et al., 2018; van der Waa et al., 2018). Contrastive explanations do not only meet the expectations of users, but are also favorable from a computational point of view: While stating the causes for a decision is outright impossible for opaque models (Doran et al., 2018), computing a contrast to a different outcome is indeed possible in most cases (van der Waa et al., 2018; Wachter et al., 2018). However, while explanations aim to assist users in interacting with AI systems, XAI research is often criticized for not putting the user in the center of attention (Abdul et al., 2018; Kirsch, 2018). Indeed, XAI research risks to experience a phenomenon referred to as "inmates running the asylum", where researchers build explanations for themselves rather than their intended users (Miller et al., 2017; Mittelstadt et al., 2019). A human-centric and "trustworthy" AI requires explanations that are understandable to the user (Gunning, 2017; European Commission, 2019). To this end, insights from social sciences might inform the design of human-centered XAI methods (Miller, 2019; Mittelstadt et al., 2019). With this paper, we answer the call for empirical studies in this context.

3 Theoretical Background

3.1 Explanations in Artificial Intelligence

The advent of deep learning enabled enormous advances in AI in the 2010s (Goodfellow et al., 2016) and resulted in the widespread application of AI systems in virtually all areas of business and everyday

life. The increasing need to provide explanations to users of such systems led to the appearance of Explainable Artificial Intelligence (XAI) as a research discipline (Gunning, 2017; Kuang, 2017). As a growing body of approaches dedicated towards making AI systems more explainable is reported in literature, the terms "interpretable", "comprehensible", and "explainable" are often used interchangeably (Chakraborty et al., 2017; Lipton, 2018). We follow the taxonomy proposed by Doran et al. (2018), which is shared in current research (e.g., Bennetot et al., 2019) to distinguish between "opaque", "interpretable", "comprehensible", and "explainable" AI systems.

The distinction between opaque and interpretable AI systems lies in the observability of their inner workings: *Opaque systems*, also known as "black boxes", do not allow the user access to the mechanism by which inputs to an AI system are mapped to its outputs (Doran et al., 2018) and further do not provide any reasoning along with the output (Caruana et al., 1999; Goodman and Flaxman, 2017). The most prominent example for opaque systems are deep neural networks (Kuang, 2017), which consist of a stack of layers of artificial neurons, whose outputs depend (usually non-linearly) on the output values of the neurons in the next-lower layer (Goodfellow et al., 2016). *Interpretable systems*, in contrast, give the user access to the mathematical description of the mapping between inputs and outputs (Doran et al., 2018). Examples include logistic regression and decision trees (Craven and Shavlik, 1999). An interpretable system might become opaque if complexity increases, for instance a regression model with many parameters. Thus, both opaque systems and interpretable systems with a high degree of complexity are a major subject of XAI research (Mittelstadt et al., 2019).

Comprehensible systems produce textual or visual symbols in addition to their output that help the user understand the mapping from the AI system's inputs to its outputs (Gedikli et al., 2014; Doran et al., 2018). One example is the family of algorithms pioneered by Ribeiro et al. (2016) with "LIME" and unified in "SHAP" by Lundberg and Lee (2017), which for any given machine learning model can determine the influence each of the inputs had on the resulting output. For instance, when detecting a hand-written digit – classified as an "8" by an opaque model – a comprehensible system could indicate which pixels in the image contributed most to the classification. *Explainable systems* go a step further: In contrast to comprehensible systems, whose output is generally only accessible to users with insight into the underlying algorithmic structure, explainable systems directly produce a human-understandable line of reasoning for why a given input is mapped to a specific output (Chakraborty et al., 2017; Abdul et al., 2018). In the example of a hand-written "8", an explainable system could provide the textual explanation: "It's an 8 because it has two circles on top of each other". Literature suggests "post hoc interpretability" (Lipton, 2018, p. 6), i.e. applying a dedicated algorithm to an opaque system, turning it into a comprehensible or explainable one (Singh et al., 2016; Doran et al., 2018; Guidotti et al., 2019). These algorithms can be model-agnostic in that they can be used for any kind of AI system and at the same time do not influence its performance (Ribeiro et al., 2016; Adadi and Berrada, 2018).

Explainable systems should produce explanations that are similar to explanations humans give to one another (Miller, 2019). These are largely constructed in a contrastive manner: People do not explain why a certain event occurred, i.e. list all its causes, but focus on why a certain event occurred instead of another, similarly perceivable one (Lipton, 1990). The basic elements of a contrastive explanation are the fact, the event that did occur, and the foil, the event that did not (Lipton, 1990; Miller, 2019). In the case of a credit rejection, the fact refers to the customer's situation (e.g., income and savings) bringing about the rejection and the foil refers to another counterfactual situation that would lead to an approval of the new line of credit (e.g., higher income). The difference between the fact and the foil is the contrast (e.g., difference in income) which is used to explain the outcome. The choice of a suitable foil is crucial for an explanation to be perceived as meaningful (Lipton, 1990; Miller, 2019). In general, the following process is conducted to automatically generate a contrastive explanation (Guidotti et al., 2018; van der Waa et al., 2018; Wachter et al., 2018): We start with a model $f(x)$, i.e. an AI system, which for a given input x_0 produces an output y_{fact} (e.g., credit line rejection). Whilst y_{fact} is determined by the model itself, an arbitrary alternative outcome y_{foil} (e.g., credit line approval) can be determined or provided by the user. Given the model $f(x)$, the input x_0, the output y_{fact}, and the alternative outcome y_{foil}, a counterfactual x is found such that $f(x) = y_{foil}$. In general, there are up to an infinite number of x for

which $f(x) = y_{foil}$ is fulfilled. Since the resulting explanation is based on the contrast $\Delta x = x - x_0$ (such that $f(x_0 + \Delta x) = y_{foil}$), the choice of x determines the characteristics of the explanation.

XAI literature provides two main lines of approaches – both of them model-agnostic – for finding an appropriate x for which $f(x) = y_{foil}$ is fulfilled and to generate a contrastive explanation: approaches relying on locally approximating the model $f(x)$ with an interpretable model from which explanations are derived (Ribeiro et al., 2016; Guidotti et al., 2018; van der Waa et al., 2018) and algorithms computing the explanation by directly accessing the inputs and outputs of the model (Wachter et al., 2018).

As major representative of the first class we introduce the Local Foil Tree approach (van der Waa et al. 2018). The foil tree can be seen as a rule extractor for opaque models (Craven and Shavlik, 1999). The approach consists of locally approximating the model around the specific output to be explained with a decision tree and deriving a contrastive explanation from its branch and leaf structure (van der Waa et al., 2018). Decision trees are considered a prime example of an interpretable model (Caruana et al., 1999; Singh et al., 2016), as the output for a given input x is the result of a chain of easily understood rules of the type "If input feature x_i is smaller than threshold t, continue with rule A, else go to rule B" (Breiman et al., 1984; Huysmans et al., 2011; Frost and Hinton, 2018). In detail, computing an explanation using a foil tree consists of five steps (van der Waa et al., 2018): First, a dataset D around x_0 is generated which includes samples classified as c_{fact} and samples classified as c_{foil} by the model $f(x)$, thereby capturing the difference between samples from these two classes. Second, a decision tree classifier is fitted to this dataset, locally approximating $f(x)$. Third, the decision path for x_0 – resulting in classification as c_{fact} – is determined from the fitted decision tree. Fourth, by assigning a weight to each decision path in the tree resulting in classification as c_{foil} and choosing the path with the lowest weight, one counterfactual decision path is selected. Finally, by calculating the difference between the two decision paths, the contrast Δx is obtained.

The second class of approaches frames the search for a suitable counterfactual x as an optimization problem where the conditions and properties of x are expressed through an objective function (Wachter et al., 2018). While a counterfactual x has to necessarily satisfy the condition that $f(x) = y_{foil}$, a contrastive counterfactual further requires that the distance $d(x_0, x)$ between x_0 and x is small and the difference $\Delta = x_0 - x$ is sparse, i.e. the factual and the counterfactual scenario should not differ too much. Based on these properties, Wachter et al. (2018) propose to find a counterfactual x to x_0 by minimizing the following objective function:

$$o(x) = \lambda \left\| f(x) - y_{foil} \right\|^2 + \sum_i \frac{|x_{0,i} - x_i|}{MAD_i} \quad (1)$$

The first term of the objective function is the squared Euclidean distance between the model's output $f(x)$ and the foil y_{foil}, weighted by a pre-factor λ. The term's minimum at $\left\| f(x) - y_{foil} \right\|^2 = 0$ is reached exactly if $f(x) = y_{foil}$, satisfying the necessary condition. The second term is a distance function which ensures that $x_0 - x$ is sparse. Wachter et al. (2018) found that the Manhattan distance $|x_0 - x|$ weighted by the mean absolute deviation MAD of each feature in the model's training dataset is most suitable in this context. In detail, finding a contrastive counterfactual x to a given x_0 consists of three steps (Wachter et al., 2018): First, the optimization problem is initialized with a given x_0 and y_{foil}, the pre-computed MAD, and a starting point $x_{initial}$. Second, a suitable optimization algorithm minimizes $o(x)$ to obtain a counterfactual x. Finally, the contrast Δx between x_0 and x is computed.

To sum up, contrastive explanations are the dominant type of human-human explanations and thus a suitable concept for XAI. The two main lines of model-agnostic approaches frame generating contrastive explanations either as an optimization problem or locally approximate the – potentially opaque – model with an interpretable one from which an explanation can be derived. In both cases, the outcome is a contrast Δx which constitutes the explanation. Although both approaches have been successfully demonstrated, to the best of our knowledge, neither has been evaluated from the users' perspective.

3.2 Evaluation of explanation methods

Recent works show promising concepts for XAI evaluation (Doshi-Velez and Kim, 2017; Abdul et al., 2018; Gilpin et al., 2018). Doshi-Velez and Kim (2017) distinguish three scenarios for the evaluation of explainable systems with varying levels of human involvement, into which existing efforts can be classified. The first scenario, *functionally-grounded evaluation*, does not require involvement of human subjects. Functionally-grounded evaluation is appropriate for methods that have already been validated via experiments with human involvement or methods that are not yet mature. For instance, Adebayo et al. (2018) tested whether saliency methods – i.e. highlighting input features relevant for the output of an AI system – are indeed model-agnostic. The second scenario, *application-grounded evaluation*, is conducted with real users in a real application setting. In this scenario, XAI methods are evaluated regarding their intended use and effect on humans, an approach that is common in human-computer interaction research (Abdul et al., 2018). For instance, Ribeiro et al. (2018) show that anchors – representing local conditions for a prediction and thereby exposing the behavior of an AI system – enable users to anticipate behavior of an AI system with less effort and higher precision. The third scenario, *human-grounded evaluation*, is conducted with human subjects in experiments as well, but on a simplified task. Human-grounded evaluation is appropriate if the focus is the quality of the explanations generated by XAI methods (Doshi-Velez and Kim, 2017; Mohseni and Ragan, 2018; Weerts et al., 2019). Indeed, it is humans who are best at evaluating how well an explanation matches human expectations (Gilpin et al., 2018). This type of experiment is common in human-computer interaction research and typically conducted with laypeople, for instance, students or workers on platforms such as Amazon's "Mechanical Turk" where users perform small tasks in exchange for a fee (Ye and Johnson, 1995; Wang et al., 2016).

Up to now, only 5% of all papers in the context of XAI include evaluation and quantification of the quality of XAI methods (Adadi and Berrada, 2018). Within this body of literature, functionally-grounded evaluation dominates with technical aspects in the center of research attention. However, XAI methods require evaluation by their most influential group of stakeholders: their users (Preece et al., 2018). This is why many authors express the need to focus more on human-grounded studies rather than just on technical aspects in order to obtain explanations enabling humans to "appropriately trust" AI systems (Wang et al., 2016; Abdul et al., 2018; Kirsch, 2018; Schneider and Handali, 2019).

3.3 Research gap

XAI is an emerging research field and key to the deployment of AI systems in practice. The need for explanations for AI systems has attracted the attention of researchers developing XAI methods (Schneider and Handali, 2019). In particular, contrastive explanations are seen as both promising and practical in building automatically generated explanations (Lim et al., 2009; Dhurandhar et al., 2018; Hoffman et al., 2018; van der Waa et al., 2018). Although XAI aims to assist users in interacting with AI systems, current research is criticized for not putting the users in the center of attention (Miller et al., 2017; Kirsch, 2018; Mittelstadt et al., 2019). In particular, there is a lack of human-grounded evaluation of existing XAI methods (Doshi-Velez and Kim, 2017). We aim to fill this gap with a human-based study that incorporates insights from social sciences in which we observe users interacting with an AI system and explanations generated by representative XAI approaches. To this end, in our paper we address the following research question: Which characteristics of explanations do users of XAI appreciate? We contribute to IS literature by deriving a set of decisive characteristics of explanations in the context of XAI and providing guidance for the design and development of user-centric XAI methods.

4 Study Design

We propose a study design to evaluate explanations generated by XAI methods from the users' perspective and identify characteristics of explanations users appreciate in the context of XAI. The basic design of the study, which is presented as the evaluation of a smartphone app for plant species detection, is as follows (cf. Figure 1): In the first step, participants are asked to match a leaf to a plant species. In the second step, they are presented the AI system's prediction and a pair of explanations referring to this

outcome. The participants are asked to choose the explanation they appreciate more. In the third step, participants justify their choice by selecting one or several reasons from a pre-defined list of explanation characteristics. The list of characteristics is derived from insights into explanations from social sciences serving as a starting point and complemented with characteristics identified in a pre-study. Participants repeat this cycle multiple times with different samples and explanations.

In the following, we describe the study and its theoretical foundations in detail. First, we derive characteristics of explanations humans appreciate when interacting with other humans. Second, we introduce the use case and the experimental setup. Subsequently, we outline the study procedure from a participant's point of view in detail. Finally, we present the XAI methods employed in the study.

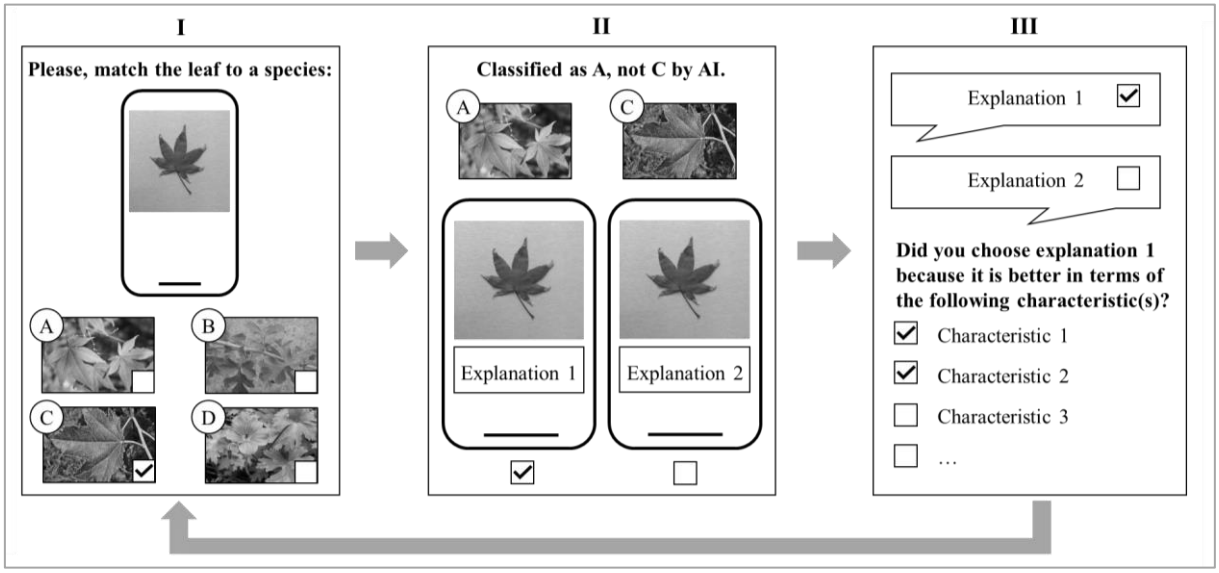


Figure 1. Basic cycle of the study: I) The participant matches the leaf shown to one of four presented species. II) The participant chooses one out of two alternative explanations for the AI system's output. III) The participant gives the reasons for their choice.

4.1 Analyzing characteristics of explanations

Interest in how humans construct and perceive explanations has generated rich literature in social sciences. In line with the definition for XAI, i.e. reasoning why a given input is mapped to an output, an explanation in human-human interaction can be defined as an exchange of causal information about an event between actors (Lewis, 1986; Lipton, 1990). Social sciences find that an explanation's "loveliness" contributes to its "likeliness" (Lipton, 2000), thus certain characteristics of an explanation beyond factual correctness can contribute to user appreciation. We take this as a starting point and derive insights on characteristics of explanations from social sciences, to then analyze these in the context of XAI.

First, explanations need to be *short*, quantified as a low "number of causes invoked in an explanation" (Lombrozo, 2007, p. 232), as humans generally tend to prefer short explanations (Thagard, 1989; Read and Marcus-Newhall, 1993; Lombrozo, 2007). However, a preference for longer explanations could be identified for scientific explanations (Weisberg et al., 2015) and in settings where a competing explanation invoked less unexplained causes (Pacer and Lombrozo, 2017). Second, explanations need to be *coherent*, i.e. relate to prior beliefs of their recipients and be overall consistent (Thagard, 1989; Lombrozo, 2012). For instance, studies found that applying prior beliefs to explain increases the perceived value of those beliefs (Preston and Epley, 2005) and consistent explanations to have greater effect on decisions (Pennington and Hastie, 1992). Third, humans prefer explanations that are *general*, i.e. explain more events (Thagard, 1989; Lombrozo, 2012). Indeed, studies found that humans tend to prefer explanations that account for more observations (Read and Marcus-Newhall, 1993) or include root causes (Kim and Keil, 2003). In contrast, a preference for narrower explanations was found when humans had to evaluate an uncertain situation with incomplete information (Khemlani et al., 2011). Fourth,

explanations need to be *relevant* to the subject of interest (Hilton and Erb, 1996; McClure, 2002). Causes are attributed higher explanatory relevance when they refer to situations that are not too far in the past (Miller and Gunasegaram, 1990), surprising (Hilton and Slugoski, 1986), or abnormal (Kahneman and Tversky, 1981; McCloy and Byrne, 2000).

To sum up, *shortness*, *coherence*, *generality*, and *relevance* are the key characteristics of explanations in human-human interaction (Thagard, 1989). These insights from social sciences can serve as a starting point when designing XAI methods (Miller, 2019). However, explanations are inherently social (Hilton, 1990) and humans are known to act differently when their counterpart is not human (e.g., Rzepka and Berger, 2018). Indeed, studies note differences in users' conversational behavior when interacting with a computer, for example regarding emotional responses (Rosenthal-von der Pütten et al., 2014) or message length (Mou and Xu, 2017). These facts lead us to the assumption that insights from social sciences require human-grounded evaluation in the context of XAI before being transferred.

4.2 Experimental setup and use case

In our experimental setup we place the user in the center of research attention, since users can best evaluate how well an explanation matches their expectations (Gilpin et al., 2018). In line with Doshi-Velez and Kim (2017), we propose a human-grounded evaluation, i.e. users interacting with XAI on a simplified task, to assess the quality of the explanations from the users' perspective. We chose our use case – an AI-based smartphone app for plant species detection – according to three guidelines for human-grounded evaluation. First, to ensure rigor, the use case should be based on real-world data and a functionally complex AI system (Abdul et al., 2018). Predictions for our app are generated by a neural network which is trained on a dataset of shape and texture attributes extracted from 340 images of leaf specimen from 30 different plant species (Silva, 2013; Silva et al., 2014). Second, the simplified task should be derived based on the needs of real-world tasks and the performance of the XAI approaches with respect to functional proxies should reflect their performance in real-world settings (Doshi-Velez and Kim, 2017). Our AI is used for classification, one of the major tasks of AI (Whittaker et al., 2018). Furthermore, our XAI methods generating contrastive explanations are model-agnostic and thus transferable to any other AI system. Third, as human-grounded evaluation is often conducted with a non-expert audience, the scenario needs to be accessible for laypeople (Doshi-Velez and Kim, 2017), while at the same time the AI system's output should not be obvious to the participant, rendering the explanations superfluous. On the one hand, we ask participants for their botanical knowledge and exclude those with expert knowledge from the study. On the other hand, only features that non-expert participants can comprehend visually (Silva et al., 2013) are included and verified as accessible in a pre-study.

4.3 Study procedure and participants

The study was conducted as a fully computerized experiment presented via a web interface. Sessions were implemented using the open-source software oTree (Chen et al., 2016) and were held online. The study procedure from a participant's point of view included an introduction and the main experiment. During the introduction, the participant read a short welcome message and responded to questions on basic demographic information (age, gender, education) as well as questions regarding potential background knowledge in AI, presence of dyschromatopsia, and botany expertise. The latter two variables later served to identify participants with expert knowledge regarding the task and those who might be limited in evaluating the explanations, respectively. Subsequently, the participant read a short description of the use case and the upcoming tasks. In the main experiment, the participant completed multiple rounds according to Figure 1. Each round started with the display of the picture of a leaf specimen, which the participant was asked to match to its plant species. To this end, four possible choices – the most probable species according to the AI's prediction – were presented along with pictures of the entire plants. We intended that the participant sometimes was in line with the AI and sometimes not – as is the case in most real-world applications. Afterwards, the participant was asked to select one out of two different contrastive explanations in a binary choice experiment (Doshi-Velez and Kim, 2017). The foil class was the participant's prediction, if they had matched wrongly, and the second most likely class

according to the predicting AI, if they had matched correctly. While selecting the better explanation, the participant was again supported by pictures of the entire plants of both the fact and the foil class. We additionally provided the opportunity to tick a box called "Both explanations are unsuitable", as there is no "guarantee that automated explanations will produce positive impact" (Ye and Johnson, 1995, p. 158) on the user. The participant was then asked to justify their selection with characteristics from a predefined list (cf. Section 5.1) and/or type a text in a box labeled "other reasons". Each participant completed 14 rounds according to Figure 1, thus judging 14 pairs of explanations. The explanations were generated by four methods (cf. Section 4.4). To avoid bias in favor of any of these methods, the participant was asked to evaluate explanations of each possible pair of methods in a random order. We supplemented the pairs of generated explanations with two pairs that each contained one false human-made explanation, i.e. an explanation which contradicts the presented picture of the leaf, which helped us identify participants who gave flippant answers (Oppenheimer et al., 2009).

We conducted a pre-study following the same procedure with 38 students to extend the list of characteristics of explanations by analyzing the "other reasons" given. Beyond, we collected feedback to improve comprehensibility of the study. We particularly revised and complemented the wording of the pre-defined list of characteristics of explanations (cf. Section 5.1) to take into account different understandings of these constructs revealed by the participants' feedback and from their free-text "other reasons". We conducted the main study with 164 participants recruited via the platform Clickworker. As is common for online survey platforms, our study further attracted 36 so-called speeders (Ford, 2017) that randomly completed the survey in conspicuously short time and were hence excluded right away. Each participant received a financial compensation for completion of the study. We chose the Clickworker marketplace since similar data quality and results can be expected compared to traditional methods and at the same time threats to internal validity are reduced (Paolacci et al., 2010; Buhrmester et al., 2011; Mason and Suri, 2012). We did not have to exclude participants due to expert botany knowledge or dyschromatopsia, but identified 20 participants who gave contradictory answers and excluded them from the study. The remaining 144 participants were equally distributed in gender (53% males, 47% females) and between 19 and 65 years old. Every participant had completed at least primary education (3% primary school, 18% middle school, 32% secondary school and 46% tertiary education). Only 17% of participants reported to be in touch with AI for work or education.

4.4 Contrastive explanation methods employed in our study

We chose two representative XAI methods generating contrastive explanations for our study (cf. Section 3.1): The Local Foil Tree (FOILTREE) as a representative of methods that generate explanations for opaque models from a local interpretable model (van der Waa et al., 2018) and an optimization algorithm (OPTIMIZE) which generates contrastive explanations directly from the opaque model (Wachter et al. 2018). To validate that the perceived difference in explanation quality is indeed due to the algorithms, we further employed the naïve approach of sampling a counterfactual from the training data (Caruana et al., 1999; Wexler, 2018). Additionally, we included human-made explanations as benchmarks (e.g., Jiang et al., 2011). In the following, the instantiations of these methods are described.

First, we instantiated FOILTREE in a standard configuration (Robeer, 2018; van der Waa et al., 2018) to our model and dataset. We generate the local training dataset for the decision tree by sampling 10,000 points along the line between the fact x_0 and a randomly chosen sample from the foil class. Further, we restrict the depth of the tree to 60 in order to obtain a generalizing model and bound the length of the resulting explanations. The algorithm outputs a vector of ranges as the contrast Δx. Second, we instantiated OPTIMIZE according to Wachter et al. (2018). Starting from a randomly sampled point in the vicinity of x_0 we find a vector x which is classified as y_{foil} with at least 60% probability from which we compute the contrast $\Delta x = x_{foil} - x_0$. To make the contrast sparser, Wachter et al. (2018) suggest to remove small values from the contrast Δx, which we implement in line with Lundberg et al. (2017) by pruning away small feature contrasts which do not significantly contribute to the classification of x as a foil. Third, we used the simple method for generating a contrast Δx of selecting an x from the training dataset for which $f(x) = y_{foil}$ and which has minimal Euclidean distance to x_0 (Caruana et al.,

1999). For the XAI methods and the naïve approach we transferred the contrast vectors $\Delta x = x - x_0$ to natural language text via a custom basic text generation engine. The resulting explanations follow the pattern "The leaf was classified as y_{fact} and not y_{foil}. In order to be classified as y_{foil}, the leaf would need to be <comparative> <adjective> … and <comparative> <adjective>." including one comparative/adjective pair for each non-zero entry of the contrast Δx. The comprehensibility of the generated texts and the adjectives, which are linked one-to-one to the features in the leaf dataset, were validated in the pre-study. Fourth, to generate human-made explanations, we asked researchers to give contrastive explanations for a random selection of leaf specimen, based on the features contained in the dataset and written with the same vocabulary and phrasing as the explanations resulting from the text generation engine.

5 Results

Our study aims to identify characteristics of explanations that users appreciate in the context of XAI. After presenting our findings, we deepen these insights by analyzing the co-occurence of the perceived length with the perception of other decisive characteristics. Finally, we analyze if and how users perceive the explanations generated by the employed XAI methods differently. In the main study, 144 participants each completed 12 rounds (excluding control rounds) evaluating pairs of explanations and justifying their choice by selecting decisive characteristics, which resulted in 1,728 evaluated pairs of explanations. We excluded pairs if participants had expressed that both explanations seemed unsuitable, leaving 1,440 pairs of explanations for further analysis. Prior to the analyses, we verified that the choice of explanations was not considerably influenced by factors other than the XAI methods' output and explanations' characteristics and that participants stayed engaged over the course of the experiment.

5.1 Identifying decisive characteristics of XAI explanations

Prior to the main study, we conducted a pre-study with 38 participants (cf. Section 4.3) that served to verify and, if necessary, expand the list of decisive characteristics of explanations derived from social sciences (cf. Section 4.1). We complemented the characteristics *short*, *coherent*, *general*, and *relevant* with *long*, i.e. the opposite of short, *concrete*, i.e. the opposite of general, and *consistent*, i.e. not containing contradictory causes, and used these seven characteristics for the main study. In the main study, participants gave one to six characteristics as reasons for their choice of explanation, on average 1.7. We aggregated the decisive characteristics for all pairs of explanations (cf. Figure 2) and conducted an exact binomial test to determine if the share of a characteristic being decisive in all pairs of explanations was significantly greater than the average share of 24.8% expected in the case of equally distributed relevance of characteristics (Gravetter and Wallnau, 2012). The share of the characteristic *concrete*, which was selected as decisive for 34.7% of all pairs of explanations is significantly greater than average ($p<0.001$), followed by *relevant* (34.3%, $p<0.001$) and *coherent* (32.9%, $p<0.001$). Other reasons to justify the choice of explanations (2.0%) given in free-text form by participants did not reveal further characteristics and hence were not considered in the analysis.

Characteristic	Number selected	Share in all rounds
concrete	499	34.7%***
relevant	494	34.3%***
coherent	474	32.9%***
long	376	26.1%
short	276	19.2%
consistent	203	14.1%***
general	182	12.6%***
average share (---)	*** p<0.001	** p<0.01 *p<0.05

Figure 2. Share of decisive characteristics for all pairs of explanations.

5.2 Analyzing co-occurrence of characteristics with perceived length

Length is discussed as a key characteristic of explanations in literature and can be measured objectively (Lombrozo, 2007; van der Waa et al., 2018; Wachter et al., 2018). Indeed, in our experiment, length (i.e. number of invoked features) significantly correlates with the probability that participants justified their selected explanation with the characteristics *long*/*short* (Pearson correlation for long: 0.73, $p<0.05$; Pearson correlation for short: -0.87, $p<0.01$). We analyze the perceived length of an explanation in more detail by investigating its co-occurrence with other characteristics. We calculate the co-occurrence of characteristics with *long* and *short*, i.e. the share of *concrete*, *relevant*, *general*, *coherent*, and *consistent* being decisive if either *short* or *long* was decisive as well. We determine 95%-confidence intervals for these shares based on Wilson score intervals (Brown et al., 2001). The analysis reveals (cf. Figure 3) that *concrete* has a considerably higher co-occurrence with *long* (33.5%, confidence interval: 28.9%-38.4%) than with *short* (19.2%, confidence interval: 15.0%-24.3%); the confidence intervals do not overlap. We conclude that users rather found an explanation to be *concrete* if they also appreciated it as *long* instead of *short*. Given that perceived and measured length are in line, we may further conclude that longer explanations tend to be appreciated as *concrete* by users in our study. Further, the analysis reveals that users rather perceived an explanation as *relevant*, *general*, or *consistent* if they also appreciated it as *short*. For the characteristic *coherent* we do not find significantly different co-occurrence with *long* or *short*.

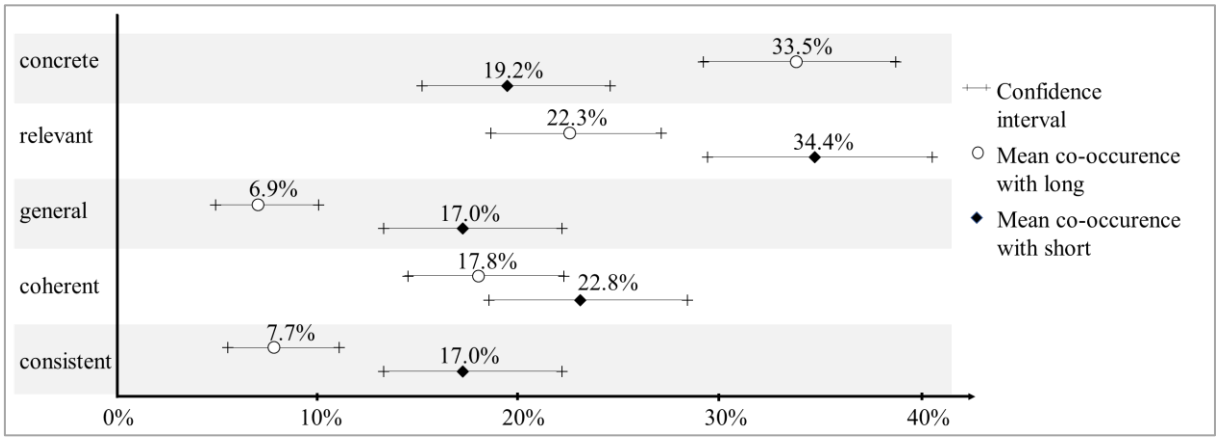


Figure 3. Co-occurrences of characteristics with long and short (95%-confidence intervals).

5.3 Evaluating contrastive explanations generated by XAI methods

Finally, we analyze how explanations generated by FOILTREE and OPTIMIZE are perceived by users with respect to the decisive characteristics. Similar to Section 5.1, we aggregate the characteristics given when a FOILTREE (OPTIMIZE) explanation was chosen as well as the characteristics given when another explanation was chosen over one generated by FOILTREE (OPTIMIZE) (cf. Figure 4). Results for explanations generated by FOILTREE are in line with the findings for all pairs of explanations, i.e. the three decisive characteristics (cf. Section 5.1) are decisive here as well. The characteristics *concrete* (34.7%, $p<0.001$), *coherent* (34.0%, $p<0.001$), and *relevant* (33.1%, $p<0.001$) were significantly more often used to justify choosing a FOILTREE explanation than the average share expected if each of the characteristics were selected with equal probability (23.8%). At the same time, the characteristics *concrete* (37.0%, $p<0.001$), *relevant* (37.0%, $p<0.001$), and *coherent* (36.3%, $p<0.001$) were significantly more often than average (24.6%) used to justify choosing the competing explanation. In contrast, for explanations generated by OPTIMIZE we find that the characteristics *concrete* (40.4%, $p<0.001$), *coherent* (36.4%, $p<0.001$), *relevant* (33.5%, $p<0.001$), and *long* (32.4%, $p<0.001$) were significantly more often than average (25.0%) used to justify choosing the competing explanation. Only the characteristic relevant (29.0%, $p<0.01$) was significantly more often than average (21.9%) used to justify choosing an OPTIMIZE explanation.

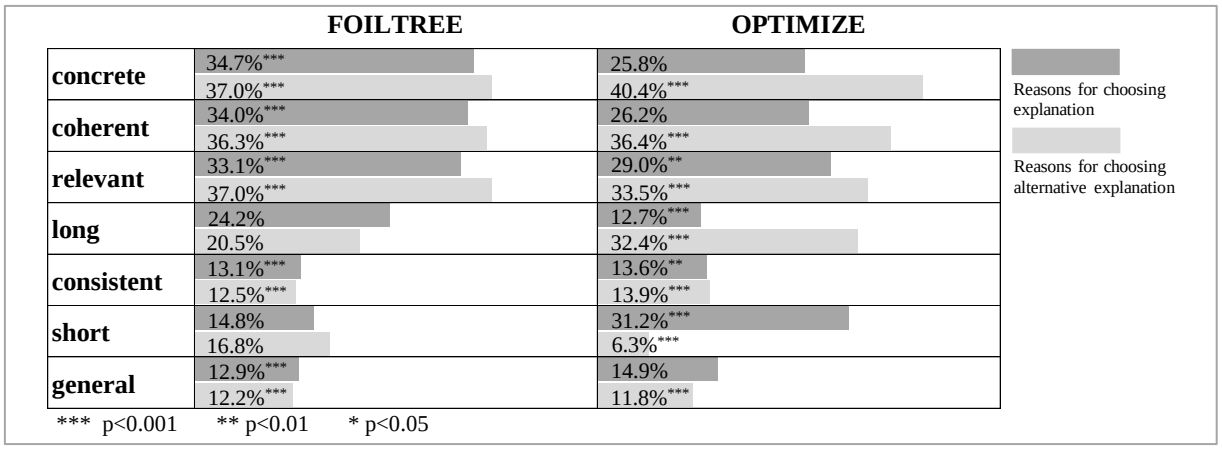


Figure 4. Share of decisive characteristics for pairs of explanations generated by XAI methods.

6 Discussion

Our study is motivated by the need for a human-centered and "trustworthy" AI (European Commission, 2019). As one step towards this aim, it is necessary to provide users with explanations that they appreciate and thus can use to make an informed decision about the recommendations of an AI system. In our study, we placed the users in the center of research attention to uncover which characteristics of explanations they appreciate when interacting with AI systems. To this end, we conducted a human-based study employing representative XAI methods and informed by insights from social sciences regarding the characteristics of explanations humans appreciate. Our results uncovered three decisive characteristics of explanations in the context of XAI: *concreteness*, *coherence*, and *relevance*. Moreover, we analyzed in depth how perceived length is co-occurring with other characteristics. Finally, we evaluated explanations generated by representative XAI methods with respect to the decisive characteristics.

6.1 Implications for theory and practice

To the best of our knowledge, we are the first putting the user in the center of research attention in order to identify decisive characteristics for automatically generated explanations for AI systems. This research contributes to a human-centric AI, as our research focuses on what characteristics users find useful when judging recommendations provided by an AI system.

First, our findings show that explanations that are *concrete*, *coherent*, and *relevant* are most appreciated by XAI users. Our findings differ from results in social sciences, which suggest *general* (Thagard, 1989; Read and Marcus-Newhall, 1993) and *short* (Thagard, 1989; Lombrozo, 2007) explanations to be appreciated most. Notably, we found that XAI users appreciate *concreteness* (decisive in 34.7% of all evaluated explanations) more than *generality* (decisive in 12.6% of all evaluated explanations). Studies suggesting that humans prefer explanations with a narrow latent scope, i.e. explanations that account for fewer unobserved effects (Khemlani et al., 2011), may underpin this result. Furthermore, we observed that users appreciate *longer* (decisive in 26.1% of all evaluated pairs of explanations) rather than *shorter* (decisive in 19.2% of pairs) explanations. Indeed, an explanation that is too short may fail to incorporate the most important aspects to sufficiently limit unexplained causes. As humans show higher level of uncertainty interacting with AI systems than when interacting with humans (Mou and Xu, 2017), unexplained causes might gain importance. Our findings are in line with insights from social sciences regarding the role of *coherence*, i.e. consistence with prior knowledge (Thagard, 1989; Lombrozo, 2012) or additional evidence provided (e.g., as in our study setting, pictures), and *relevance* to the subject of interest (Hilton and Erb, 1996; McClure, 2002). We thus support the call that insights into explanations found in human-human interaction should inform the design of XAI methods (Miller et al., 2017; Abdul et al., 2018) but argue that they need to be carefully evaluated before being transferred.

Second, we provide guidance for the design and development of user-centric XAI methods that help to foster human agency by providing users with the arguments needed to contest or follow the recommendation of an AI system. Our finding that *concreteness*, *coherence*, and *relevance* are the decisive characteristics of explanations from the users' perspective provides an empirically grounded target for XAI research. Literature has so far focused on the need for *general* explanations (van der Waa et al., 2018; Miller, 2019) and *short* explanations (Narayanan et al., 2018; Miller, 2019). While *coherence* was already identified as highly important (Miller, 2019), to the best of our knowledge no XAI method explicitly incorporates *coherence* in explanation generation. Still, XAI research has made progress in focusing on the most *relevant* causes (Mittelstadt et al., 2019), with some XAI methods starting to explicitly address this characteristic (van der Waa et al., 2018). From a theoretical perspective, we call for research in the design of XAI methods that addresses the so far predominantly neglected characteristics *coherence* and *concreteness*. Practical implications can be derived for the implementation of XAI methods with respect to the characteristic *length*. We found that if *length* was appreciated in an explanation, also *concreteness* was rather appreciated. Further, if *shortness* was appreciated, *relevance*, *generality*, and *consistency* were more likely to be appreciated. These findings suggest that XAI methods should employ explanation length in a strategic manner, i.e. reducing length, but not at any expense. While prior research mainly focuses on short explanations, our results indicate an explanation that is too short may fail to convince users of its *concreteness*. However, as *shortness* is linked to *relevance*, our findings suggest that a concentration on few but decisive explanatory causes can increase the users' appreciation. Thus, XAI methods should in practice be calibrated to produce explanations with a length that allows for sufficient depth, but still restricts the explanation to the most relevant causes. From these general conclusions, we can derive potential improvements for the major representative approaches generating contrastive explanations employed in our study. Our results suggest that users perceive the characteristics of explanations generated by FOILTREE as varying: The reasons to choose an explanation generated by FOILTREE correspond with the characteristics overall appreciated by XAI users. However, these characteristics were also the most frequently given reasons for choosing an alternative explanation. Thus, we conclude that FOILTREE is generally able to generate excellent explanations from the users' perspective and in a next step would benefit from stabilizing its output. In contrast, the decisive characteristics (except for *relevance*) were not notably appreciated by users when they preferred an explanation generated by OPTIMIZE, but instead given when choosing an alternative explanation. In particular, in 36.4% of cases users preferred the alternative explanation because it was *coherent*, while only in 26.2% of cases users appreciated this characteristic when choosing explanations by OPTIMIZE. A similar tendency can be observed for *concreteness*, which was the reason users preferred the alternative explanation in 40.4% of cases, which is in line with our finding that it co-occurs with *length*. Thus, calibrating length might be a first adjustment to improve OPTIMIZE to better address the decisive characteristics.

Third, we contribute to theory by providing a generic study design to evaluate explanations generated by XAI methods. Such evaluation is needed to guide researchers and developers in building explanations that empower the user when interacting with an AI system. The methodological considerations which informed the design of our human-based study for evaluating XAI methods proved appropriate and thus may serve as reference point for further research. Both the use case and the procedure of the study – asking the participants to make a decision, choose between two alternative explanations, and justify their choice – were successful in uncovering which traits of explanations users appreciate. The short cycles at the core of the study allowed to present a large number of explanations to each participant, while the generic nature of the procedure allows for transfer to other use cases. As intended, the chosen use case was accessible to laypeople, but sufficiently challenging as to not make the matching of leaves to the respective plant species or the explanations obvious for the participants. Holding the study online and recruiting participants on the Clickworker platform resulted in a large population diverse with respect to age, educational background, and gender. Finally, the pre-study proved valuable not only to identify a set of decisive characteristics of explanations, but also to fine-tune comprehensibility and validate the design and viability of the entire experiment. With our evaluation of explanations generated by two major representative XAI methods we hope to encourage other researchers to evaluate their XAI methods with users.

6.2 Limitations and further research

Our study is subject to several limitations. First, due to the focus on one specific use case and in absence of similar comparable studies, the external validity is necessarily limited. While the use case was modelled to exhibit traits found in many real-world AI applications, it is nevertheless artificial. However, its suitability for laypeople and the fact that the dataset is publicly available render it predestined to serve as the testbed for future research. Second, our study focused on contrastive explanations, based on the insight that these are the dominant kind of explanations in human-human interaction, and specifically two major representative methods for their generation. However, various other XAI methods and ways to present explanations exist. Contrastive explanations alone can be presented in various ways, e.g., visually, as tabular data, or with different variants of natural language text. As their presentation may significantly influence how explanations are perceived (e.g., Huysmans et al., 2011), this harbors a large, yet mostly untapped potential for furthering XAI explanations. Our study design is well suited to explore this. Third, the participants were not directly affected by or personally invested in the AI system's decision. It is to be expected that the characteristics of explanations users appreciate differs in scenarios where users need to act based on an explained AI decision (e.g., AI assistants in a professional context) or are subject to it (e.g., credit approval). Future studies might address this point by tasking users to act based on an AI system's output. Fourth, limited cognitive abilities of participants – which is inherent in human-based experiments – might reduce the internal validity of the results. Although we believe human-grounded evaluation to be most appropriate to extract decisive characteristics for automatically generated explanations for AI systems, further experiments might validate and expand on our results. Fifth, while laypeople are an important group amongst users of AI systems, it is also necessary to understand what explanations are appreciated by expert users (e.g., medical professionals). It is expected that the characteristics of explanations experts appreciate differ from the preferences of lay users and also vary amongst different professions. Overall, the study presented in this paper constitutes a first but important step towards user-centric XAI and sets the stage for future research. As a next step, the results obtained can be used to further improve and develop XAI methods with a focus on decisive characteristics from the users' perspective. In this context, measuring and quantifying these characteristics bears great potential to guide the design and development without the constant need for costly and time-consuming human evaluation. While the characteristic *length* already has direct, objective measure, this is not the case for the characteristics *concreteness*, *coherence*, and *relevance*. Finally, future studies might go beyond appreciation of explanations and place a focus on the effectiveness of explanations in the context of human-AI interaction by investigating use cases where the user is subject to or can base their next action on an AI decision. In this context, long-term effects, where over time users get accustomed to explanations, seem especially of interest with regard to real-world applications.

7 Conclusion

Considerable progress has been made in developing automatically generated explanations for AI systems. However, XAI research is criticized for not putting the user in the center of attention. In this paper, we developed a human-based study to evaluate explanations generated by XAI methods from the users' perspective. We took insights from social sciences as a starting point to derive characteristics of explanations XAI users appreciate. We conducted a user study with 164 participants evaluating contrastive explanations generated by major representative XAI approaches. Our results revealed *concreteness*, *coherence*, and *relevance* as decisive characteristics. We further found that XAI users rather find an explanation to be *concrete* if they also appreciate it as *long* and *relevant*, *general*, or *consistent* if they also appreciate it as *short*. Finally, we identified potential improvements of XAI methods generating contrastive explanations. Our contribution to IS literature and practice is threefold: We are the first to derive and validate a set of decisive characteristics of explanations in the context of XAI. Second, we provide guidance for the user-centric development of XAI methods. Third, our work provides a generic study setup for the evaluation of XAI methods from the users' perspective. We hope to encourage other researchers to evaluate XAI methods with users, complementing the progress in the design of XAI methods, thereby pushing the fascinating research field of XAI forward.

References

Abdul, A., J. Vermeulen, D. Wang, B. Y. Lim and M. Kankanhalli (2018). "Trends and Trajectories for Explainable, Accountable and Intelligible Systems." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal, QC: ACM Press.

Adadi, A. and M. Berrada (2018). "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6, 52138–52160.

Adebayo, J., J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt and B. Kim (2018). "Sanity Checks for Saliency Maps." In: *Advances in Neural Information Processing Systems 31*, pp. 9505–9515.

Angelino, E., N. Larus-Stone, D. Alabi, M. Seltzer and C. Rudin (2018). "Learning Certifiably Optimal Rule Lists for Categorical Data." *Journal of Machine Learning Research* 18 (234), 1–78.

Benbasat, I. and W. Wang (2005). "Trust In and Adoption of Online Recommendation Agents." *Journal of the Association for Information Systems* 6 (3), 72–101.

Bennetot, A., J.-L. Laurent, R. Chatila and N. Díaz-Rodríguez (2019). "Towards Explainable Neural-Symbolic Visual Reasoning." In: *Proceedings of the 2019 International Workshop on Neural-Symbolic Learning and Reasoning*. Macao, pp. 71–75.

Bloomberg, J. (2018). *Don't Trust Artificial Intelligence? Time To Open The AI "Black Box."* Forbes Enterprise & Cloud. URL: https://www.forbes.com/sites/jasonbloomberg/2018/09/16/dont-trust-artificial-intelligence-time-to-open-the-ai-black-box/ (visited on 10/31/2019)

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone (1984). *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall/CRC.

Brown, L. D., T. T. Cai and A. DasGupta (2001). "Interval Estimation for a Binomial Proportion." *Statistical Science* 16 (2), 101–117.

Bughin, J., E. Hazan, S. Ramaswamy, M. Chui, T. Allas, P. Dahlström, N. Henke and M. Trench (2017). *Artificial Intelligence: The next digital frontier?* McKinsey Global Institute.

Buhrmester, M., T. Kwang and S. D. Gosling (2011). "Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science* 6 (1), 3–5.

Caruana, R., H. Kangarloo, J. D. N. Dionisio, U. Sinha and D. B. Johnson (1999). "Case-based explanation of non-case-based learning methods." In: *AMIA 1999, American Medical Informatics Association Annual Symposium*. Washington, DC: AMIA, pp. 212–215.

Chakraborty, S., R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis and P. Gurram (2017). "Interpretability of deep learning models: A survey of results." In: *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*. San Francisco, CA: IEEE.

Chen, D. L., M. Schonger and C. Wickens (2016). "oTree—An open-source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance* 9, 88–97.

Chui, M., S. Francisco and J. Manyika (2018). *Notes from the AI Frontier: Insights from Hundreds of Cases*. McKinsey Global Institute.

Craven, M. W. and J. W. Shavlik (1999). "Rule Extraction: Where Do We Go From Here?" *University of Wisconsin Machine Learning Research Group Working Papers* (99–1).

Dhurandhar, A., P. Chen, R. Luss, C. Tu, P. Ting, K. Shanmugam and P. Das (2018). "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives." In: *Advances in Neural Information Processing Systems 31*, pp. 592–603.

Doran, D., S. Schulz and T. R. Besold (2018). "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives." In: *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*. Bari.

Doshi-Velez, F. and B. Kim (2017). "Towards A Rigorous Science of Interpretable Machine Learning." *ArXiv* 1702.08608.

European Commission (2019). *Ethics Guidelines for Trustworthy AI*. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 (visited on 03/16/2020)

Ford, John B. (2017). "Amazon's Mechanical Turk: A Comment." *Journal of Advertising* 46(1), 156–

158
Freedy, A., E. DeVisser, G. Weltman and N. Coeyman (2007). "Measurement of trust in human-robot collaboration." In: *Proceedings of the 2007 International Symposium on Collaborative Technologies and Systems*. Orlando, FL: IEEE, pp. 106–114.
Frost, N. and G. Hinton (2018). "Distilling a neural network into a soft decision tree." In: *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*. Bari.
Gedikli, F., D. Jannach and M. Ge (2014). "How should I explain? A comparison of different explanation types for recommender systems." *International Journal of Human-Computer Studies* 72 (4), 367–382.
Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter and L. Kagal (2018). "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning." In: *The 5th IEEE International Conference on Data Science and Advanced Analytics*. Turin: IEEE.
Goodfellow, I., Y. Bengio and A. Courville (2016). *Deep Learning*. Cambridge, MA: MIT Press.
Goodman, B. and S. Flaxman (2017). "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation."" *AI Magazine* 38 (3), 50–57.
Gravetter, F. J. and L. B. Wallnau (2012). *Statistics for the Behavioral Sciences*. 9th Edition. Belmont, CA: Wadsworth Publishing.
Guidotti, R., A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini and F. Giannotti (2018). "Local Rule-Based Explanations of Black Box Decision Systems." *ArXiv* 1805.10820.
Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi and F. Giannotti (2019). "A Survey Of Methods For Explaining Black Box Models." *ACM Computing Surveys* 51 (5).
Gunning, D. (2017). *Explainable Artificial Intelligence (XAI)*. DARPA. URL: https://www.darpa.mil/program/explainable-artificial-intelligence (visited on 11/25/2019)
Herse, S., J. Vitale, M. Tonkin, D. Ebrahimian, S. Ojha, B. Johnston, W. Judge and M. Williams (2018). "Do You Trust Me , Blindly? Factors Influencing Trust Towards a Robot Recommender System." In: *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication*. Nanjing: IEEE.
Hilton, D. J. (1990). "Conversational processes and causal explanation." *Psychological Bulletin* 107 (1), 65–81.
Hilton, D. J. and H.-P. Erb (1996). "Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance." *Thinking & Reasoning* 2 (4), 273–308.
Hilton, D. J. and B. R. Slugoski (1986). "Knowledge-Based Causal Attribution. The Abnormal Conditions Focus Model." *Psychological Review* 93 (1), 75–88.
Hoffman, R., T. Miller, S. T. Mueller, G. Klein and W. J. Clancey (2018). "Explaining Explanation, Part 4: A Deep Dive on Deep Nets." *IEEE Intelligent Systems* 33 (3), 87–95.
Huysmans, J., K. Dejaeger, C. Mues, J. Vanthienen and B. Baesens (2011). "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models." *Decision Support Systems* 51 (1), 141–154.
Jiang, Y. G., G. Ye, S. F. Chang, D. Ellis and A. C. Loui (2011). "Consumer video understanding: A benchmark database and an evaluation of human and machine performance." In: *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. Trento: ACM Press.
Kahneman, D. and A. Tversky (1981). *The Simulation Heuristic*. Office of Naval Research.
Khemlani, S. S., A. B. Sussman and D. M. Oppenheimer (2011). "Harry Potter and the sorcerer's scope: Latent scope biases in explanatory reasoning." *Memory and Cognition* 39 (3), 527–535.
Kim, N. and F. Keil (2003). "From symptoms to causes: Diversity effects in diagnostic reasoning." *Memory & Cognition* 31 (1), 155–165.
Kirsch, A. (2018). "Explain to whom? Putting the user in the center of explainable AI." In: *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*. Bari.
Kuang, C. (2017). *Can A.I. Be Taught to Explain Itself?* The New York Times Magazine. URL: https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html (visited on 11/26/2019)
Lee, M. K. (2018). "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management." *Big Data & Society* 5 (1).

Lewis, D. K. (1986). "Causal explanation." *Philosophical Papers* 2, 214–240.
Lim, B. Y., A. K. Dey and D. Avrahami (2009). "Why and why not explanations improve the intelligibility of context-aware intelligent systems." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Boston, MA: ACM Press, pp. 2119–2128.
Lipton, P. (1990). "Contrastive Explanation." *Royal Institute of Philosophy Supplement* 27, 247–266.
Lipton, P. (2000). "Inference to the Best Explanation." In: W. H. Newton-Smith (Ed.), *A Companion to the Philosophy of Science*. Maiden, MA: Blackwell.
Lipton, Z. C. (2018). "The Mythos of Model Interpretability." *Queue* 16 (3), 1–27.
Lombrozo, T. (2007). "Simplicity and probability in causal explanation." *Cognitive Psychology* 55 (3), 232–257.
Lombrozo, T. (2012). "Explanation and Abductive Inference." In: K. J. Holyoak & R. G. Morrisson (Eds.), *The Oxford Handbook of Thinking and Reasoning*. Oxford: Oxford University Press.
Lorica, B. and P. Nathan (2018). *The State of Machine Learning Adoption in the Enterprise*. 1st Edition. (M. Slocum, Ed.). Sebastopol, CA: O'Reilly.
Lundberg, S. and S.-I. Lee (2017). "A Unified Approach to Interpreting Model Predictions." In: *Advances in Neural Information Processing Systems 30*. Long Beach, CA.
Mason, W. and S. Suri (2012). "Conducting behavioral research on Amazon's Mechanical Turk." *Behavior Research Methods* 44 (1), 1–23.
McCloy, R. and R. M. J. Byrne (2000). "Counterfactual thinking about controllable events." *Memory & Cognition* 28 (6), 1071–1078.
McClure, J. (2002). "Goal-based Explanations of Actions and Outcomes." *European Review of Social Psychology* 12 (1), 201–235.
Miller, D. T. and S. Gunasegaram (1990). "Temporal Order and the Perceived Mutability of Events: Implications for Blame Assignment." *Journal of Personality and Social Psychology* 59 (6), 1111–1118.
Miller, T. (2019). "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence* 267, 1–38.
Miller, T., P. Howe and L. Sonenberg (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-17 Workshop on Explainable AI (XAI)*. Melbourne, pp. 36–42.
Mittelstadt, B., C. Russell and S. Wachter (2019). "Explaining explanations in AI." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Atlanta, GA: ACM Press, pp. 279–288.
Mohseni, S. and E. D. Ragan (2018). "A Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning." *ArXiv* 1801.05075.
Mou, Y. and K. Xu (2017). "The media inequality: Comparing the initial human-human and human-AI social interactions." *Computers in Human Behavior* 72, 432–440.
Narayanan, M., E. Chen, J. He, B. Kim, S. Gershman and F. Doshi-Velez (2018). "How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation." *ArXiv* 1802.00682.
Oppenheimer, D. M., T. Meyvis and N. Davidenko (2009). "Instructional manipulation checks: Detecting satisficing to increase statistical power." *Journal of Experimental Social Psychology* 45 (4), 867–872.
Pacer, M. and T. Lombrozo (2017). "Ockham's Razor cuts to the root: Simplicity in causal explanation." *Journal of Experimental Psychology: General* 146 (12), 1761–1780.
Paolacci, G., J. Chandler and P. G. Ipeirotis (2010). "Running experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5 (5), 411–419.
Pennington, N. and R. Hastie (1992). "Explaining the Evidence: Tests of the Story Model for Juror Decision Making." *Journal of Personality and Social Psychology* 62 (2), 189–206.
Preece, A., D. Harborne, D. Braines, R. Tomsett and S. Chakraborty (2018). "Stakeholders in Explainable AI." In: *AAAI FSS-18: Artificial Intelligence in Government and Public Sector Proceedings*. Arlington, VA.
Preston, J. and N. Epley (2005). "Explanations Versus Applications: The Explanatory Power of

Valuable Beliefs." *Psychological Science* 16 (10), 826–832.
Rader, E. and R. Gray (2015). "Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul: ACM Press, pp. 173–182.
Ransbotham, S., D. Kiron, P. Gerbert and M. Reeves (2017). "Reshaping Business With Artificial Intelligence: Closing the Gap Between Ambition and Action." *MIT Sloan Management Review*.
Read, S. J. and A. Marcus-Newhall (1993). "Explanatory coherence in social explanations: A parallel distributed processing account." *Journal of Personality and Social Psychology* 65 (3), 429–447.
Ribeiro, M. T., S. Singh and C. Guestrin (2016). ""Why Should I Trust You?"" In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM Press, pp. 1135–1144.
Ribeiro, M. T., S. Singh and C. Guestrin (2018). "Anchors: High-precision model-agnostic explanations." In: *32nd AAAI Conference on Artificial Intelligence*. New Orleans, LA: Association for the Advancement of Artificial Intelligence, pp. 1527–1535.
Robeer, M. J. (2018). *Contrastive Explanation for Machine Learning*. Utrecht University. URL: https://dspace.library.uu.nl/handle/1874/368081 (visited on 11/28/2019)
Rosenthal-von der Pütten, A. M., F. P. Schulte, S. C. Eimler, S. Sobieraj, L. Hoffmann, S. Maderwald, M. Brand and N. C. Krämer (2014). "Investigations on empathy towards humans and robots using fMRI." *Computers in Human Behavior* 33, 201–212.
Rzepka, C. and B. Berger (2018). "User Interaction with AI-enabled Systems: A systematic review of IS research." In: *Proceedings of the 39th International Conference on Information Systems*. San Francisco, CA: Association for Information Systems.
Schneider, J. and J. P. Handali (2019). "Personalized Explanation for Machine Learning: A Conceptualization." In: *Proceedings of the 27th European Conference on Information Systems*. Stockholm/Uppsala.
Silva, P. F. B. (2013). *Development of a System for Automatic Plant Species Recognition*. Universidade do Porto. URL: https://repositorio-aberto.up.pt/handle/10216/67734 (visited on 11/28/2019)
Silva, P. F. B., A. R. S. Marcal and R. A. da Silva (2014). *Leaf Dataset*. UCI Machine Learning Repository. URL: https://archive.ics.uci.edu/ml/datasets/Leaf (visited on 12/06/2018)
Silva, P. F. B., A. R. S. Marçal and R. M. A. da Silva (2013). "Evaluation of Features for Leaf Discrimination." In: *International Conference Image Analysis and Recognition*. Póvoa do Varzim: Springer, pp. 197–204.
Singh, S., M. T. Ribeiro and C. Guestrin (2016). "Programs as Black-Box Explanations." In: *Proceedings of NIPS 2016 Workshop on Interpretable Machine Learning for Complex Systems*. Barcelona.
Thagard, P. (1989). "Explanatory coherence." *Behavioral and Brain Sciences* 12, 435–502.
van der Waa, J., M. Robeer, J. van Diggelen, M. Brinkhuis and M. Neerincx (2018). "Contrastive Explanations with Local Foil Trees." In: *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning*. Stockholm, pp. 41–46.
Wachter, S., B. Mittelstadt and C. Russell (2018). "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law & Technology* 31 (2), 841–887.
Wang, N., D. V. Pynadath and S. G. Hill (2016). "Trust calibration within a human-robot team: Comparing automatically generated explanations." In: *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. Christchurch: IEEE Press, pp. 109–116.
Weerts, H. J. P., W. van Ipenburg and M. Pechenizkiy (2019). "A Human-Grounded Evaluation of SHAP for Alert Processing." *ArXiv* 1907.03324.
Weisberg, D. S., J. C. V Taylor and E. J. Hopkins (2015). "Deconstructing the seductive allure of neuroscience explanations." *Judgment and Decision Making* 10 (5), 429–441.
Wexler, J. (2018). *The What-If Tool: Code-Free Probing of Machine Learning Models*. Google AI Blog. URL: https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html (visited on 11/24/2019)
Whittaker, M., K. Crawford, R. Dobbe, G. Fried, E. Kaziunas, V. Mathur, S. M. West, R. Richardson,

J. Schultz and O. Schwartz (2018). *AI Now Report 2018*. New York, NY: AI Now Institute.
Ye, L. R. and P. E. Johnson (1995). "The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice." *MIS Quarterly* 19 (2), 157–172.

3.2 Fostering Human Agency: A Process for the Design of User-Centric XAI Systems

Full Citation:	Förster, Maximilian; Klier, Mathias; Kluge, Kilian; & Sigler, Irina (2020). Fostering Human Agency: A Process for the Design of User-Centric XAI Systems. In *Proceedings of the 41st International Conference on Information Systems*, Virtual Conference, 1-17. https://aisel.aisnet.org/icis2020/hci_artintel/hci_artintel/12/
Copyright Note:	Reprinted according to author's rights.

Dec 14th, 12:00 AM

Fostering Human Agency: A Process for the Design of User-Centric XAI Systems

Maximilian Förster
University of Ulm, maximilian.foerster@uni-ulm.de

Mathias Klier
University of Ulm, mathias.klier@uni-ulm.de

Kilian Kluge
University of Ulm, kilian.kluge@uni-ulm.de

Irina Sigler
University of Ulm, irina.hardt@uni-ulm.de

Fostering Human Agency: A Process for the Design of User-Centric XAI Systems

Completed Research Paper

Maximilian Förster
University of Ulm
Helmholtzstr. 22
89081 Ulm, Germany
maximilian.foerster@uni-ulm.de

Mathias Klier
University of Ulm
Helmholtzstr. 22
89081 Ulm, Germany
mathias.klier@uni-ulm.de

Kilian Kluge
University of Ulm
Helmholtzstr. 22
89081 Ulm, Germany
kilian.kluge@uni-ulm.de

Irina Sigler
University of Ulm
Helmholtzstr. 22
89081 Ulm, Germany
irina.sigler@uni-ulm.de

Abstract

The emerging research field of Explainable Artificial Intelligence (XAI) addresses the problem that users do not trust or blindly follow AI systems that act as black boxes. XAI research to date is often criticized for not putting the user at the center of attention. Against this background, we design a process to systematically guide the instantiation, calibration, and quality control of XAI systems such that they foster human agency and enable appropriate trust in AI systems. The process can be applied independent of the XAI method, application domain, and target user group. It incorporates the principles of user-centric design, insights into explanations from the social sciences, and established XAI evaluation scenarios. Following the Design Science methodology, we demonstrate the practical applicability of our artifact and evaluate its efficacy in a realistic setting. Our work contributes to the design of user-centric XAI systems and the quest for human agency in AI.

Keywords: Explainable Artificial Intelligence, User-Centric Design, Human-AI Interaction

Introduction

Artificial Intelligence (AI) is increasingly employed for a wide range of tasks, first and foremost, decision support (HLEG-AI 2019). However, "Without AI systems [..] being demonstrably worthy of trust, unwanted consequences may ensue and their uptake might be hindered" (HLEG-AI 2019, p. 4). Thus, AI systems need to guarantee human agency (HLEG-AI 2019), as otherwise, users might either blindly follow an AI system's recommendation or merely distrust and not use it (Herse et al. 2018; Rader and Gray 2015). The key impediment to human agency is the fact that many AI systems appear as "black boxes" that do not provide users with sufficient information to make an informed choice regarding their recommendations (Guidotti et al. 2019b; Wachter et al. 2018).

In light of this challenge, the research field of Explainable Artificial Intelligence (XAI) aims at AI systems that are both highly performant and empower their users to comprehend, appropriately trust, and scrutinize them (Abdul et al. 2018; DARPA 2017). In particular, XAI provides approaches to automatically generate explanations along with AI systems' outputs (Rai 2020). In this context, explanations are human-

understandable lines of reasoning for why an AI system maps a given input to a specific output (Abdul et al. 2018). As the primary motivation for providing explanations is to enable human agency (HLEG-AI 2019; Nunes and Jannach 2017), the user-centricity of explanations is a prerequisite (Ribera and Lapedriza 2019). User-centricity is the "extent to which a system, product, or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO 9241-210 2019, sec. 3.13). However, while substantial progress has been made in developing and demonstrating XAI methods (Barredo Arrieta et al. 2020), research to date is criticized for not putting the users at the center of attention (Kirsch 2018; Mittelstadt et al. 2019). Recent research has begun to address this call by examining insights from social sciences (Miller 2019) and evaluating explanations from users' perspectives (Doshi-Velez and Kim 2018; Förster et al. 2020; Weerts et al. 2019). However, while these efforts yield first valuable insights into a better understanding of XAI users, findings remain fragmented, and users are still not systematically incorporated into the development of XAI methods. This creates a situation best described as "inmates running the asylum," with researchers constructing explanations they themselves appreciate rather than explanations that generate value for their users (Miller et al. 2017; Mittelstadt et al. 2019). Indeed, to the best of our knowledge, no approach exists that effectively and systematically guides researchers and practitioners in the user-centric design of XAI systems.

Against this background, we propose a novel IT artifact that guides researchers and practitioners in instantiating, calibrating, and controlling the quality of user-centric XAI systems. The design of our artifact is informed by prior work on user-centric design, insights into understanding XAI users, and methods to evaluate XAI systems. Our artifact takes the shape of a process inspired by well-established processes in the fields of data mining and data science. Following the Design Science methodology (Hevner et al. 2004; Sonnenberg and vom Brocke 2012), we demonstrate and rigorously evaluate our process by applying it to a use case transferable to other AI applications. Our contribution to research and practice is twofold. First, we conceptualize and evaluate a user-centric XAI process to guide researchers and practitioners in the design of XAI systems. Second, we demonstrate how to effectively incorporate processes from data mining and data science, principles of user-centric design, insights from the social sciences into characteristics, structures, and presentation modes of explanations, and evaluation frameworks in XAI into a unified process.

The remainder of this paper is structured as follows: In the next section, we discuss relevant literature in the fields of user-centricity, XAI, as well as data mining and data science that inform the design of our artifact. Subsequently, we propose a process for instantiating, calibrating, and controlling the quality of a user-centric XAI system. Then, we demonstrate the applicability of the artifact and evaluate its efficacy. Afterward, we discuss the implications of our research for theory and practice, reflect on limitations of our work, and conclude with directions for further research.

Theoretical Background

Explanations for AI decisions

A significant issue of many state-of-the-art AI systems is their opacity, or "black box" character, which means that their inner workings are so intricate that the reasons for their decisions appear impenetrable to the user (Guidotti et al. 2019b). A prominent example of opaque systems are deep neural networks (Doran et al. 2018), which are comprised of a stack of layers of artificial neurons, whose outputs depend (typically non-linearly) on the outputs of the neurons in the next-lower layer (Goodfellow et al. 2016). Opacity induces critical challenges regarding the adoption of AI systems and resulting consequences. First, opacity hinders AI's societal acceptance, as it contributes to users' distrust in the AI's decisions and consequently reduces their willingness to consider or accept recommendations (Herse et al. 2018). Second, opacity impedes human agency, as users lack the information and transparency needed to reflect critically on an AI system's decision before following or acting on it (Rader and Gray 2015). Explanations that accompany the AI system's decisions can provide the level of transparency needed to scrutinize AI decisions (HLEG-AI 2019; Nunes and Jannach 2017), enabling users to appropriately trust the system (DARPA 2017; HLEG-AI 2019). Accordingly, they are seen as a promising path in the quest for a trustworthy AI (HLEG-AI 2019).

In light of the challenges of both low AI adoption due to a lack of users' trust and the harmful consequences of AI systems that impede human agency, the research field of XAI provides algorithms for automatically generating explanations (Doran et al. 2018). The call for explanations for AI systems has attracted

considerable attention from researchers. For an overview, see the reviews by Barredo Arrieta et al. (2020) and Guidotti et al. (2019b). XAI systems, in their most basic form, consist of an algorithm generating explanations for an AI system and an explanation interface (DARPA 2017). Often, XAI algorithms are model-agnostic (Rai 2020, Guidotti et al. 2019b) and generate "post hoc interpretations" (Lipton 2018, p. 6). Thus, they can be used for any kind of AI system while not influencing its performance.

XAI explanations build on elements such as visualizations, feature-relevance, or counter-examples, with most approaches using a combination thereof. Visualizations convey the reasons for a decision through plots and graphics, e.g., partial dependence plots (Green and Kern 2010). Feature-relevance explanations measure the importance each feature has in generating the output, e.g., through estimation of Shapley values (Lundberg and Lee 2017). Algorithms generating counter-examples go a step further and explain a decision by contrasting it to another comparable decision (Wachter et al. 2018), inspired by how humans construct explanations themselves (Lipton 2000). XAI literature suggests two main lines of approaches to finding a suitable counter-example such that the explanation is meaningful: algorithms relying on locally approximating the AI system with a simpler model from which explanations are derived (Guidotti et al. 2019a) and algorithms computing explanations directly from the AI system, often framing the search for a counter-example as an optimization problem (Dhurandhar et al. 2019; Wachter et al. 2018). While the technical realization of XAI methods is an essential prerequisite, the call for explanations goes beyond providing post hoc interpretation for an AI system's output (Miller et al. 2017; Mittelstadt et al. 2019). It requires solutions that not only explain the recommendations of an AI system to users who do not understand its inner workings but further enable these users to contest and alter a recommendation (Doran et al. 2018; Wachter et al. 2018). In light of these challenges, to fully support human agency, the research field of XAI needs to place its users at the center of attention (Abdul et al. 2018; Kirsch 2018).

User-Centric XAI

In the context of human agency, individuals are seen as "contributors to their life circumstances, not just products of them" (Bandura 2006). In line with this definition, the Independent High-Level Expert Group on AI set up by the European Commission demands that AI systems guarantee human agency, as "users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system." (HLEG-AI 2019, p. 16). Explanations are a crucial element of allowing for such informed decisions, as they aim to enable the subject to understand the reasons for a decision as well as put them into a position to contest or affect it (Wachter et al. 2018). In this line of thought, empowering users to control and appropriately trust an AI system is the primary motivation for providing explanations (HLEG-AI 2019; Nunes and Jannach 2017). Thus user-centricity of explanations is a prerequisite (Ribera and Lapedriza 2019). Still, whereas "researchers in the ML and AI communities are working on making their algorithms explainable, their focus is not on usable, practical and effective transparency that works for and benefits people" (Abdul et al. 2018, p. 10). Indeed, while XAI research provides a wide array of algorithms to produce a diverse range of explanations for AI recommendations, it remains unclear what the end-user needs to scrutinize and appropriately trust an AI system (Förster et al. 2020; Wang et al. 2019).

User-centric design might answer this call, as it provides a design approach to developing solutions that focus on the users' needs and wants (Norman and Draper 1986). The establishment of user-centric design in the 1980s (cf. Norman and Draper 1986) marks a milestone in product and service development (Still and Crane 2017). In general, user-centric design is an approach that aims at improving usability, namely the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO 9241-210 2019, sec. 3.13). In the context of XAI, this can be transferred to an approach that puts XAI users, whether they are laypeople or domain experts, at the center of attention and enables them to achieve the goal of empowering users to control and appropriately trust an AI system (Ribera and Lapedriza 2019). The premise of user-centric design has inspired a broad range of methods and principles (cf. Still and Crane 2017). With their fundamental principles, Gould and Lewis (1985) proposed an early focus on users and tasks, empirical measurement, and iterative design as crucial elements of user-centric design. The guidelines put forward by IDEO, a leading design-agency, additionally identify empathy and the need to learn from failure as vital principles (IDEO 2015). Likewise, guidelines such as the "People + AI Guidebook" by Google (2019) that specifically focus on human-centric AI products emphasize the need to consider user-centricity throughout

the entire product development flow and provide guidance on, e.g., identifying user needs and the design of feedback mechanisms. The generic ISO guideline for "human-centered design" provides a well-established framework, frequently used in academia and practice. It proposes six principles (ISO 9241-210 2019): First, the need for understanding user, task, and environment. Second, user involvement in design and development. Third, user-centric evaluation. Fourth, the need for an iterative process that, fifth, addresses the entire user experience. Sixth, a multidisciplinary team with diverse skills and perspectives is required. Based on these principles, the ISO guideline identifies four key activities that a user-centric design process needs to entail, namely, understanding the context, specifying user requirements, producing the solution, and evaluating the solution. If necessary, several iterations of these activities are to be performed until a satisfactory solution can be instantiated. Together, the principles and activities serve as general guidelines to achieve a user-centric design that can be adapted for application in specific contexts. For instance, Farinango et al. (2015) integrated them into user-centric software development processes.

Research into explanations for AI systems that represent human-understandable lines of reasoning and enable human subjects to gain control when interacting with an AI system and develop an appropriate level of trust (Abdul et al. 2018) can build on strong foundations (cf. Wang et al. 2019). Social sciences find that an explanation's "loveliness" contributes to its "likeliness" (Lipton 2000) and point out specific characteristics, structures, and presentation modes of explanations beyond factual correctness that can contribute to user appreciation. First, social sciences literature identifies explanation characteristics, e.g., shortness (Thagard 1989), that are appreciated in human-human interaction and hence might inform the design of XAI systems (cf. Förster et al. 2020; Miller 2019). Second, regarding the basic structure of an explanation, research refers to how humans construct explanations themselves, suggesting XAI methods to produce contrastive explanations (e.g., Wachter et al. 2018). These explanations do not list all causes that lead to a specific event but focus on why an AI system yielded a particular output (the *fact*) instead of another, similarly perceivable one (the *foil*) (cf. Lipton 1990). The difference between the fact and the foil, the *contrast*, explains the output. In the case of the rejection of a new credit line, the fact refers to the customer's situation (e.g., income and savings) leading to the rejection. The foil refers to a counterfactual scenario that would bring about an approval (e.g., higher income). The contrast is the difference between the customer's situation and the counterfactual scenario (e.g., difference in income). Aside from characteristics and structure, specific modes of presentation can improve intelligibility. These include, among others, visual, textual, symbolic, audible, audio-visual, or tabular (Wang et al. 2019). For example, Huysmans et al. (2011) found that decision tables are especially comprehensible, while Ribera and Lapedriza (2019) demonstrated that visualization serves to support decision-making in healthcare.

Prior work on the evaluation of automatically generated explanations provides first insights on the incorporation of user-centricity into the design of XAI systems. In this context, Doshi-Velez and Kim (2018) propose three scenarios for the evaluation of explainable systems. The first, *functionally-grounded evaluation*, does not require human involvement. One potential approach is to test against proxy measures for explanations, e.g., the length of an explanation as a measure for its simplicity or complexity, respectively (Martens and Provost 2014; Wachter et al. 2018). While efficient in terms of time and resource requirements, it remains unclear whether such proxy measures truly reflect the users' perception of explanations. Thus, the second scenario, *human-grounded evaluation*, is conducted with human subjects undertaking a simplified task to assess the quality of explanations from users' perspective (Förster et al. 2020; Mohseni and Ragan 2018; Weerts et al. 2019). The third scenario, *application-grounded evaluation* with real users in a real application setting, can serve as the final step in evaluating usability and effectiveness (Abdul et al. 2018). Practitioners and researchers are confronted with trade-offs when choosing the most suitable evaluation scenario, most notably the trade-off between including users and required effort. On the one hand, conducting experiments with human subjects is crucial for lowering the risk of being misled by assumptions that do not reflect users' perception (Weerts et al. 2019). On the other hand, both expenditure of time and costs are generally substantially higher for the human-grounded compared to the functionally-grounded scenario (Doshi-Velez and Kim 2018).

To sum up, while first valuable insights into user-centric explanations that can inform the design of XAI methods have been reported, to the best of our knowledge, no process for the systematic application of user-centric principles to the design of XAI systems exists. This situation poses a challenge for researchers and practitioners looking to incorporate user-centricity in the design of XAI methods, as they face a highly fragmented state of knowledge.

Data Mining and Data Science Processes

Incorporating principles of user-centric design into the design of systems is a challenge that arises beyond the field of XAI. In the broadest sense, the design of information systems always entails the challenge of solving technical tasks while meeting pre-defined objectives. To address this challenge, research areas related to XAI, e.g., data mining and data science, rely on processes (Martinez-Plumed et al. 2019). More specifically, these processes guide projects by translating business goals into well-defined technical tasks (Marbán et al. 2009). This general approach can be transferred to the design of XAI systems, as – similar to data and machine-learning models – their design requires an explorative approach, the translation of pre-defined business objectives into technical metrics (e.g., accuracy), and statistical testing of the solution. Thus, we take processes from the fields of data mining and data science as a starting point to incorporate user-centric objectives into the design of XAI systems. Data mining literature defines a process as a series of steps that are executed in sequence. It can include loops and iterations, which are "triggered by a revision process" (Kurgan and Musilek 2006, p. 4). Data mining processes typically contain three stages: They begin with steps to understand the business goals and context, followed by data preparation and analysis, and conclude with the evaluation, interpretation, and application of the results (Kurgan and Musilek 2006; Martinez-Plumed et al. 2019). As an example, consider CRISP-DM, the de facto standard process for data mining in research and practice (Martinez-Plumed et al. 2019). This process comprises six steps, namely, business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman et al. 2000). Many researchers based their further developed processes on CRISP-DM (Martinez-Plumed et al. 2019). For instance, Gertosio and Dussauchoy (2004) expanded the process to include increased user involvement. More recently, faced with new challenges such as ever-larger data volumes and the rise of machine learning, both IBM and Microsoft released updated and more versatile variants (Microsoft 2020; Rollins 2015). Data mining and data science processes serve to minimize risks through different validation steps, to reveal and remedy faults, and to facilitate resource allocation (Marbán et al. 2009; Rollins 2015). Due to their flexibility and scalability, they can be applied independent of project size and domain (Marbán et al. 2009). Further, the processes provide a general and replicable framework that allows projects to be executed by staff with diverse backgrounds (Moyle and Jorge 2001). Finally, the clear goal-definition enforced by data science processes fosters alignment between team members (Microsoft 2020). To sum up, processes in the areas of data mining and data science can inform the incorporation of user-centricity into the design of XAI systems. In particular, their basic structure can serve as a blueprint when designing a novel process for the design of user-centric XAI systems.

Research Gap

In order to be beneficial to individuals and society, the proliferation of AI in everyday life requires that users are able "to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system" (HLEG-AI 2019, p. 4). In this regard, especially the opacity inherent to many AI systems is an impediment, as it leads to human subjects facing AI decisions without the means required to understand and contest them. Against this background, in the quest for trustworthy AI, the emerging research field of XAI aims to provide explanations that foster human agency (HLEG-AI 2019; Nunes and Jannach 2017). Over the past years, the automatic generation of explanations received tremendous research attention that resulted in the development of a wide range of algorithms (Barredo Arrieta et al. 2020). However, a majority of these have not yet been evaluated with human users, and the application of XAI systems in real-world contexts is still in its infancy (Abdul et al. 2018; Adadi and Berrada 2018). While first studies recently addressed the call for putting the user in the center of research attention (Kirsch 2018), e.g., examining insights from social sciences (Miller 2019) and evaluating explanations from users' perspectives (Förster et al. 2020), these findings remain fragmented. To the best of our knowledge, no process exists that systematically guides the design of user-centric XAI systems. Therefore, researchers and practitioners alike are at a high risk of designing XAI systems that do not provide value for their users (Miller et al. 2017; Mittelstadt et al. 2019). To close this gap, following the Design Science methodology (Hevner et al. 2004), we design and evaluate a novel process to instantiate, calibrate, and control the quality of user-centric XAI systems. Our process focuses on model-agnostic XAI methods and the end-users of XAI systems, such as domain experts and laypeople. The process can be applied to any application domain that entails an AI system augmenting human decision-making, excluding systems that fully automate it (Martin 2019). It places the users at the center of attention while striking a balance between costly and time-consuming user testing and calibration based on mathematical constructs.

A Novel Process to Design User-Centric XAI Systems

We design a novel process to instantiate, calibrate, and control the quality of an XAI system such that it is user-centric in that it enables and fosters human agency (DARPA 2017; HLEG-AI 2019). To this end, we design our "User-Centric XAI Process" (cf. Figure 1) based on well-established processes in the field of data mining and data science (cf. Martinez-Plumed et al. 2019) and the principles of user-centric design (ISO 9241-210 2019). The process integrates the evaluation framework for explainable systems by Doshi-Velez and Kim (2018) and incorporates research on explanations in the social sciences (cf. Miller 2019).

Basic Idea

We begin by briefly revisiting the problems researchers and practitioners face when designing XAI systems. First of all, XAI methods are novel algorithms that have yet to stand the test of practice and time (Wolf 2019). Further, contrary to the task-driven development of AI systems, in the context of XAI, users should be at the center of attention (Preece et al. 2018). Importantly, user-centricity in XAI reaches far beyond usability, as human agency is the primary goal, and further legal and ethical concerns demand consideration (HLEG-AI 2019). Due to a lack of experience and best practices to draw from, real-world XAI applications are at risk of failing their users and falling short of their stakeholders' high expectations (Weerts et al. 2019). Not because XAI methods are inherently incapable – quite the converse (Barredo Arrieta et al. 2020) – but because the designers of XAI systems lack the means to shift their focus from technical aspects to their lay or domain-expert end-users (Miller et al. 2017). Our artifact addresses this problem space based on the three core concepts sequential structure, user-centricity, and iterative calibration.

First, we design our artifact to provide a *sequential structure* that systematically guides the design of XAI systems. Inspired by well-established processes in the related fields of data mining and data science (Martinez-Plumed et al. 2019), we design a process comprising the phases Instantiation, Calibration, and Quality Control (cf. Kurgan and Musilek 2006). The Instantiation phase focuses on examining the application requirements as well as selecting and instantiating the XAI system. In the subsequent Calibration phase, the XAI system is adapted to produce explanations that fulfill the application-specific requirements in an iterative sequence of calibration and user testing. Finally, in the Quality Control phase, the deployed XAI system is continuously monitored and evaluated to assess its efficacy. Designing the artifact as a process exhibits three main advantages: First, a process provides a structured and replicable framework to systematically develop complex systems. Second, the process prescribes the definition of precise and unambiguous goals and ensures that they are not lost out of sight throughout the potentially lengthy and intertwined stages of system development. Third, the process places the technical development of the XAI system in the context of its users and the team designing it.

Second, our artifact emphasizes *user-centricity* by placing the end-user at the center of attention, as arguably, the users are the most critical stakeholders of XAI systems (Preece et al. 2018). On the one hand, usability is crucial for the successful application of an XAI system, as users have to accept and interact with it. On the other hand, fostering human agency is the primary goal in the design of XAI systems. However, users are rarely considered in XAI research to date (Kirsch 2018). By incorporating the principles of user-centric design (ISO 9241-210 2019) and building on XAI literature on user-centric explanations (cf. Wang et al. 2019), our "User-centric XAI process" ensures that the user is in focus at all times. For one, the Instantiation phase fosters a thorough understanding of the task, user, and environment, ensuring that the entire user experience is taken into account from the very beginning. Both the subsequent Calibration and Quality Control phase suggest an iterative approach guided by user feedback (Gould and Lewis 1985; IDEO 2015; ISO 9241-210 2019). Further, the design process is informed by insights from the social sciences regarding characteristics and presentation modes of explanations (cf. Miller 2019; Wang et al. 2019), such as the human preference for contrastive explanations or the need for explanations to be coherent.

Third, our artifact integrates the complementary XAI evaluation scenarios proposed by Doshi-Velez and Kim (2018) into a unified process of *iterative calibration* to enable efficient yet user-centric design of XAI systems. While user testing and involvement are indispensable to ensure user-centricity (ISO 9241-210 2019; Weerts et al. 2019), it is costly and time-consuming. Hence, it cannot be carried out continuously, but only on selected occasions. Against this background, we integrate functionally-grounded and human-grounded evaluation by interlinking them with proxy measures, i.e., mathematical constructs that reflect the user-centric requirements for the XAI system (Doshi-Velez and Kim 2018). Functionally-grounded

evaluation serves to calibrate the XAI system through optimizing its parametrization to specified target values of these proxy measures. This activity requires no user involvement and is thus economical in time and costs. To rigorously validate that the proxy measures truly reflect the users' perspective, we employ human-grounded evaluation, which tests explanations with human subjects, often on a simplified task that aims to capture the essential elements and features of the application setting (Doshi-Velez and Kim 2018).

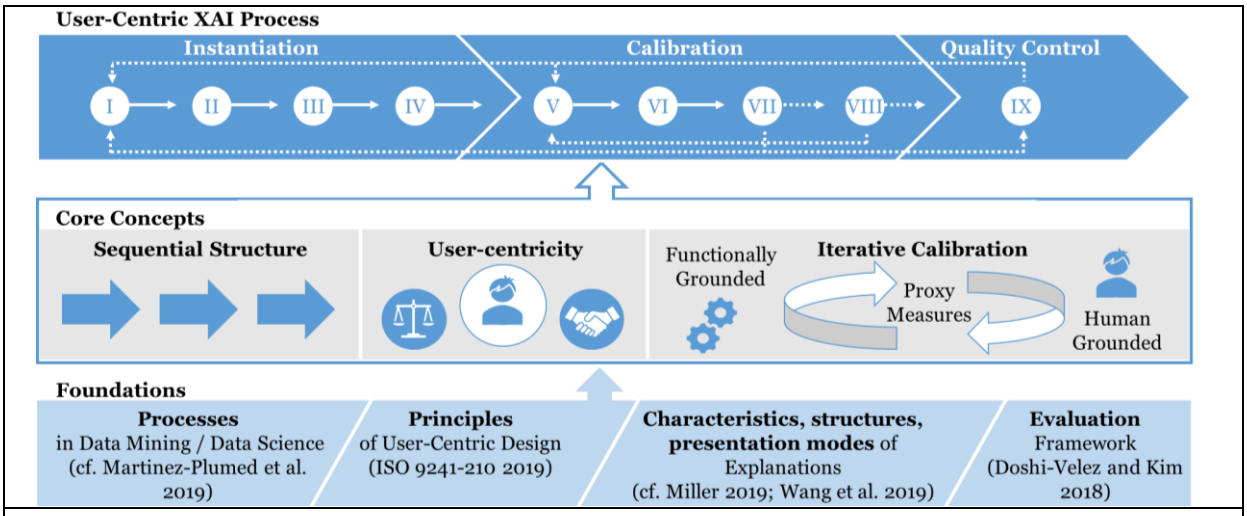


Figure 1. Basic idea and underlying core concepts of the User-Centric XAI Process

In the following, we describe the "User-Centric XAI Process" with its three phases and its nine corresponding steps (cf. Table 1) in detail. As a starting point, we assume a ready-to-use AI system that is accessible throughout the process. In line with the fifth principle of user-centric design (ISO 9241-210 2019), the process is conducted by a team with diverse backgrounds in, e.g., IS, AI, social sciences, or business-specific domains (cf. Mittelstadt et al. 2019).

Phases and Steps

The **Instantiation phase**, which consists of four steps, is devoted to understanding the XAI system's application domain and the target users as well as defining requirements for the XAI system. Step I aims to build an understanding of the specific XAI application context. First, in line with typical data mining processes (Chapman et al. 2000; Kurgan and Musilek 2006; Microsoft 2020), the team investigates the XAI system's intended application domain and the corresponding business background. More concretely, ethical, legal, and regulatory requirements, as well as the availability of resources such as data and computation infrastructure, are examined (cf. Chapman et al. 2000; Martinez-Plumed et al. 2019). Second, informed by the first principle of user-centric design, i.e., the need for understanding user, task, and environment (Gould and Lewis 1985; IDEO 2015; ISO 9241-210 2019), the level of domain expertise the user possesses, the purpose of the XAI system for the users, and the risks associated with a lack of user-centric explanations are examined (Beaudouin et al. 2020). Next to the primary goal of enabling human agency, further aims, e.g., legal requirements derived from the responsibility for consequences resulting from the AI system's use, are captured (Beaudouin et al. 2020; HLEG-AI 2019; Mittelstadt et al. 2019). Further, the team needs to establish an understanding of the role the AI system plays in the decision-making process (Google 2019). More specifically, this includes placing the intended application on the augmentation-automation continuum and understanding the respective roles of the human agent and the AI system (Martin 2019). Additionally, the team should identify the potential for decision-making bias in the intended application (cf. Baron 2014; Tversky and Kahneman 2015) and investigate whether literature already suggests potential mitigation measures in the application domain, e.g., as in the case of medical diagnosis (cf. Lighthall and Vazquez-Guillamet 2015; Wang et al. 2019). Based on the developed understanding of the XAI system's application context and purpose, in Step II, the team derives a list of user-centric and technical application requirements from the users' perspective (cf. IBM 2016; Moyle and Jorge 2001). Then, the team needs to identify whether these requirements can be translated into specific modes of presentation and desired characteristics of explanations, respectively. More concretely, technical

modes of presentation concern how the explanations are conveyed to the user and range from visual, textual, or symbolic presentation to audible, audio-visual, or interactive explanations (Wang et al. 2019). The choice of a mode of presentation is informed by the AI system's technical structure and its application context, with a focus on user experience (cf. ISO 9241-210 2019). Characteristics of explanations describe their perception by the user, e.g., concrete, coherent, or relevant (Förster et al. 2020), and reflect the intention of the explanations. At this early stage of the process, the precise characteristics required to fulfill the intentions are not yet known. Hence, the team identifies initial characteristics guided by the literature (Miller 2019) or user interviews (Hall et al. 2019). While prior XAI research indicates a variety of requirements in different user groups and settings (Hall et al. 2019; Kirsch 2018; Miller 2019; Ribera and Lapedriza 2019; Wang et al. 2019), systematically building an in-depth user understanding constitutes an innovative contribution to XAI. Equipped with user-centric and technical requirements, in Step III, the team selects an XAI method as the foundation of the XAI system. Similar to the selection of models in data mining processes (Kurgan and Musilek 2006) and data science (Goodfellow et al. 2016; Microsoft 2020), the selection is informed by the team's expertise and can require substantial literature research. In line with Hall et al. (2019), to be considered as the foundation of the XAI system, an XAI method has to be capable of generating explanations with the modes of presentation identified in Step II. Moreover, it needs to offer sufficient flexibility in its parametrization, such that its explanations can be tuned to exhibit the desired characteristics. The instantiation of the XAI system concludes Step III. At the end of the Instantiation phase, in Step IV, a set of preliminary proxy measures is derived. Proxy measures reflect the intended characteristics of explanations and are computed from the XAI system's output (Doshi-Velez and Kim 2018). The preliminary proxy measures can be based on examples from the literature (Dhurandhar et al. 2019; Guidotti et al. 2019b; Mothilal et al. 2020; Wachter et al. 2018) or, at this point in the process, be derived based on the team's intuition.

The four steps of the subsequent **Calibration phase** aim at calibrating the XAI system according to iteratively refined requirements. The alternation between functionally-grounded and human-grounded evaluation exploits the resource efficiency of the former. At the same time, the latter addresses the issue that most proxy measures available from the XAI literature have not been validated with users and are not equally applicable in all contexts (Adadi and Berrada 2018; Doshi-Velez and Kim 2018; Förster et al. 2020). The goal of Step V is to find a parametrization of the XAI system that leads to optimal explanations as indicated by the proxy measures. To systematically explore the XAI system's potentially vast parameter space, well-established methods for tuning machine-learning algorithms, such as grid search (Goodfellow et al. 2016), can be employed. In terms of the framework by Doshi-Velez and Kim (2018), each calculation of proxy measures for a possible parametrization of the XAI system constitutes a functionally-grounded evaluation. While research into XAI methods often ends with this step, generating explanations that satisfy the researchers' intuition (Miller et al. 2017; Mittelstadt et al. 2019), an explanation "is not a mathematical construct" but "the users are the leveling rule" (Kirsch 2018). Thus, to test if explanations exhibit the desired characteristics from the users' perspective, in Step VI, in accordance with the principles of user-centric design (ISO 9241-210 2019), we conduct a user test. More concretely, in a scenario of human-grounded evaluation (Doshi-Velez and Kim 2018), users evaluate explanations from the XAI system against competing and control explanations on a simplified task. Control explanations can, for instance, be ones written by experts (Förster et al. 2020) or instantiations of the XAI system from previous iterations. As far as possible, participants of the user test should be representative of the target audience. In the case of lay users, online platforms can provide access to a diverse demographic (Buhrmester et al. 2011). The user test yields data on how the explanations generated by the XAI system are perceived compared to control and competing explanations (Doshi-Velez and Kim 2018; Förster et al. 2020). Moreover, the test reveals which characteristics are most important from the users' perspective (Förster et al. 2020). In Step VII, the results of the user test are analyzed in two ways. First, the desired characteristics of explanations from the users' perspective are derived. Second, it is determined to what extent the XAI system produces explanations that exhibit these desired characteristics. This analysis constitutes a validation step, echoing the concept of validation in data science processes (Marbán et al. 2009; Rollins 2015): If the results are satisfying, the XAI system is fit for deployment, and the process can continue with the Quality Control phase. If a large fraction of explanations generated by the XAI system is found to be unsuitable, this might hint at oversights or incorrect assumptions in previous steps. In this case, the process should be discontinued and resumed with Step I. In all other cases, the results of the user test inform the adaption of the desired characteristics of explanations and the refinement of proxy measures in Step VIII. The desired characteristics are assessed and, if necessary, adapted based on both the conducted analysis and the investigation in Step I. Using the

data collected in the user test, it is analyzed whether each proxy measure indeed reflects its corresponding characteristic. If it does not, it is discarded. Subsequently, new proxy measures are introduced for characteristics that are not yet or not sufficiently captured. The refinement of the proxy measures concludes an iteration of the Calibration phase; the process proceeds with a new iteration, beginning with Step V.

The final **Quality Control phase** is devoted to continuous evaluation of the deployed XAI system under real-world conditions. First, the output of the XAI system is monitored, similar to the monitoring and maintenance of applications developed in data mining and data science processes (Microsoft 2020; Moyle and Jorge 2001). In a scenario of continuous functionally-grounded evaluation (Doshi-Velez and Kim 2018), it is observed if generated explanations fall in the identified target range of the final set of proxy measures, e.g., through a dashboard (Microsoft 2020). Second, corresponding to the application-grounded evaluation scenario described by Doshi-Velenz and Kim (2018), the effects of the XAI system on its users are observed. One potential area of investigation might be the presence of cognitive bias that impairs decision-making (cf. Baron 2014; Tversky and Kahneman 2015) and the extent to which the XAI system serves to mitigate or amplify it. An early and thorough assessment of the XAI system's impact and efficacy is especially important in cases where the participants of the user test conducted in Step VI and the target users differ significantly. Thus, e.g., in the case of a consumer entertainment application, which could be tested with a diverse group of lay users during calibration, a small-sized survey might be sufficient. On the other end of the spectrum, a system that supports decision making in healthcare may require several rounds of testing (Wang et al. 2019). If an application-grounded evaluation reveals that the XAI system does not yet or no longer meet its requirements, the process can be resumed with Step I or V at the team's discretion.

Table 1. Steps and Tasks of the User-Centric XAI Process

Instantiation	I Context and user specification	• Investigate application domain and business background • Identify the purpose of the XAI system • Create an understanding of the target user group
	II Application requirements	• Derive user-centric and technical requirements • Identify modes of presentation and desired characteristics of explanations based on requirements
	III Instantiation of XAI method	• Select an XAI method as the foundation of the XAI system • Instantiate the XAI system
	IV Preliminary proxy measures	• Select a preliminary set of proxy measures
Calibration	V Calibration with proxy measures	• Find parameters of the XAI system that lead to optimal explanations according to the proxy measures
	VI Evaluation with users	• Conduct a user test to evaluate generated explanations against competing and control explanations
	VII Analysis of users' perception	• Derive desired explanation characteristics from users' perspectives • Evaluate if explanations generated by the XAI system meet the desired explanation characteristics • *If results are satisfying, go to Step IX, if results are unsuitable, go to Step I, otherwise go to Step VIII*
	VIII Refinement of proxy measures	• Validate and refine the desired characteristics of explanations • Validate and adapt the set of proxy measures • *Go to Step V*
Quality Control	IX Evaluation and monitoring	• Continuously monitor the XAI system's output • Evaluate the XAI system's efficacy under real-world conditions • *If the XAI system does not fulfill its requirements, go to Step I or V*

Table 1. Steps and Tasks of the User-Centric XAI Process

Demonstration and Evaluation

As an essential part of the Design Science research process (Hevner et al. 2004), we demonstrate and evaluate the applicability and efficacy of our artifact in a realistic setting. We provide quantitative evidence that the artifact fulfills its objective and rigorously assess the efficacy of its core concepts.

Setting

Evaluation of the applicability and effectiveness of a design artifact requires demonstration of an instantiation in a setting that closely resembles the three "realities" system, task, and users (Sonnenberg and vom Brocke 2012). Against this background, we select an AI-based smartphone app for plant species detection as our use case. With the app, lay users can take pictures of leaves, whose species is detected by an AI system and displayed as text (cf. Förster et al. 2020). The task is to add an XAI system to the app that generates accompanying textual explanations that help the user understand the AI system's reasoning. We use a simulated prototype of the app closely modeled after real-world examples, such as the smartphone app Plantix or the AI system for identifying plant diseases presented by Ramcharan et al. (2019). The AI system classifies leaves using a neural network trained on a publicly available real-world dataset of shape and texture attributes extracted from 340 images of leaf specimen from 30 different plant species (Silva et al. 2013, 2014).

Application of the User-centric XAI Process

At the beginning of the Instantiation phase, in Step I, we identified that the app targets a lay audience with an interest in nature. In turn, explanations had to be comprehensible without expert knowledge. The focus of explanations was to entertain and educate users, helping them to gradually improve their botany knowledge while casually interacting with the app. Building on these insights, in Step II, we chose contrastive explanations, as the primary purpose of the explanations was to convey information about the leaf classification (cf. Miller 2019). Motivated by both comprehensibility and constrained smartphone screen space, we opted for short textual natural-language explanations displayed alongside the picture of the classified leaf as the mode of presentation. Turning to characteristics, in addition to the comprehensibility requirement, we assessed that explanations should be faithful to the AI system. While explanations should be as general as possible to facilitate learning, they should nevertheless fully explain the specific classification result. Research in social sciences suggests shortness, generality, and coherence (Lombrozo 2012; Thagard 1989), as well as relevance (Hilton and Erb 1996; McClure 2002), as characteristics of explanations that humans generally value. As the starting point to select an XAI method, in Step III, we considered that the AI system used for plant species detection utilized a feature extraction algorithm to extract the leaf's features from a picture (Silva et al. 2013, 2014). Thus, the input to the neural network at the heart of the AI system is a vector x_{fact} with numerical, scalar features, which could serve as the input to the XAI system as well. We selected the popular, frequently used algorithm proposed by Wachter et al. (2018), which is compatible with the input data, ensures faithfulness by directly operating on the AI system, and can be computed efficiently for neural networks (Dhurandhar et al. 2019). In a nutshell, this algorithm computes contrastive explanations by framing the search for a suitable counterfactual as an optimization problem. The approach builds on minimizing an objective function with two terms: First, the squared and weighted Euclidean distance between the AI system's output $f(x)$ and the foil y_{foil}. Second, the Manhattan distance $|x_{fact} - x|$ weighted by the mean absolute deviation MAD of each feature in a representative dataset, to ensure that $x_{fact} - x$ is sparse:

$$o(x) = \lambda \left\| f(x) - y_{foil} \right\|^2 + \sum_i \frac{\left| x_{\text{fact},i} - x_i \right|}{MAD_i} \qquad (1)$$

The XAI method by Wachter et al. (2018) has several parameters that influence the properties of generated explanations: First, the parameter λ in its objective function balances the foil's faithfulness with the sparsity of the contrast. Second, the optimization can either be conducted for a specified amount of steps or stopped once the AI system's classification of the foil surpasses a confidence threshold. Third, the resulting contrast vector can be pruned of small values. To this end, setting a threshold (Wachter et al. 2018), pruning greedily to arrive at the minimal contrast that sustains the foil's classification (Mothilal et al. 2020), or pruning features in order of ascending feature importance (Förster et al. 2020; Lundberg and Lee 2017) are all

established options. We transferred the contrast vector $\Delta x = x_{foil} - x_{fact}$ to natural language text via a custom basic text generation engine (Förster et al. 2020). The resulting explanations follow the pattern "The leaf was classified as y_{fact} and not y_{foil}. In order to be classified as y_{foil}, the leaf would need to be <comparative> <adjective> … and <comparative> <adjective>." including one comparative/adjective pair for each non-zero entry of the contrast Δx. In Step IV, we defined a set of preliminary proxy measures. Initially, we did not have insight into which characteristics were valued by the users of the plant species detection app. Hence, we selected simple measures for faithfulness, comprehensibility, and generality based on literature. To measure faithfulness, we determined whether the foil was indeed classified as the foil class (Martens and Provost 2014). As a preliminary proxy measure for comprehensibility, we determined the length of an explanation by counting the number of non-zero entries in the contrast (Wachter et al. 2018). Finally, we used the distance to the closest point in the AI system's training dataset as a preliminary proxy measure for the generality of explanations (Guidotti et al. 2019a).

We conducted three iterations of the Calibration phase. In each iteration, in Step V, we undertook a grid search of the XAI system's parameter space to find a configuration that performed well regarding the proxy measures (cf. Goodfellow et al. 2016). To this end, we selected a set of values for each parameter of the XAI system, instantiated it for each possible combination of parameter values, and generated explanations for 100 facts randomly sampled from the dataset. We calculated each proxy measure's value for each of the explanations and selected the parameter combination that best fitted the desired value range and balance. Then, in Step VI, we conducted a binary choice experiment with users in the shape of an online study presented via a web interface built with the oTree framework (Chen et al. 2016). In the experiment, users interacted with the simulated prototype of the app. First, they were presented with a leaf picture (the fact) and asked to match it to one of four possible plant species. Had the user matched correctly, the second-most likely plant-species, according to the AI system, subsequently served as the foil. Were they mistaken, the plant species they had selected was used as the foil. Second, the user saw two alternative explanations, either generated by the XAI system or through a control method, e.g., an explanation written by a researcher or generated by picking the closest data point labeled as the foil class from the dataset. Users selected the explanation they preferred or indicated when they found both to be unsuitable. Third, users selected the characteristics that influenced their decision from a pre-defined list and had the opportunity to give additional free-text justification. All users completed multiple cycles, each time judging a new pair of explanations. The study setup is described in more detail in Förster et al. (2020).

In the following, we detail each of the three iterations of the Calibration phase, with a special focus on the analysis of the user test (Step VII) and subsequent refinement of the proxy measures (Step VIII).

In the first iteration, our main focus was on tuning the XAI system such that it produced both faithful and short explanations (median length 1 with mean absolute deviation from the median (MAD) 1.0, generality 98%, faithfulness 100%). We subsequently conducted a user test with a small number of users (N=38) recruited among university students. While the audience was not representative of the intended target demographic, it allowed for a cost-efficient and rapid first validation of the application requirements. Through analysis of the collected data (Step VII), we uncovered the full set of decisive characteristics in the application context. As expected, this set comprised shortness, coherence, generality, and relevance. Additionally, it included length, concreteness, and consistency, which we found by analyzing users' free-text justifications. When assessing which characteristics were valued most, we found that short explanations consisting of just a single feature were often considered inferior to longer explanations by the users, contrary to XAI literature (Martens and Provost 2014; Wachter et al. 2018). Accordingly, in Step VIII, we relaxed the goal of creating explanations that were as short as possible. However, as the evaluation had revealed additional characteristics and we, therefore, had not gathered data on the perception of explanations regarding the complete set of characteristics, we decided to undergo another iteration and keep the set of proxy measures unaltered.

For the second iteration, we calibrated the XAI system to yield longer explanations (median length 2 with MAD 0.93, generality 98%, faithfulness 100%). We conducted a user test with significantly more users (N=144) recruited on the online platform Clickworker. This population was diverse in age, educational background, and gender and closely resembled the target audience in this regard. In Step VII, we analyzed which characteristics users named most frequently when selecting an explanation. This analysis revealed concreteness (named for 34.7% of judged pairs), coherence (34.3%), and relevance (32.9%) as the decisive characteristics of explanations, which users chose significantly more frequently than the expected average

($p<0.001$, one-sided binomial test). A subsequent analysis of the relationship between perceived characteristics uncovered a co-occurrence of concreteness and length (selected together in 33.5% of cases, Wilson score 95%-confidence interval 28.9%-38.4%) as well as a co-occurrence between relevance and shortness (34.4%, 29.1%-40.2%). In Step VIII, we observed a strong correlation between the proxy measure for length and users' perception of length (Pearson correlation 0.73, $p<0.05$) and shortness (-0.87, $p<0.01$). However, it remained unclear whether the number of features itself or linguistic properties such as the length of the sentence or the presence of comparatives were the decisive factor. Regarding the proxy measures for faithfulness and generalizability, the evaluation results did not reveal a clear link with any of the desired characteristics. We concluded that in the given scenario, users were not necessarily looking for a complete or faithful explanation, but were satisfied with a concrete explanation coherent with their expectations. While we abandoned the generalizability measure, we continued to require faithfulness, since the delivery of correct explanations was an important requirement. In summary, we found that users generally perceived the explanations generated by the XAI system as lacking in concreteness, relevance, and coherence, but sometimes appreciated their shortness. Motivated by the clear link between length and perceived characteristics, we more closely analyzed the collected data in this regard. We observed that users perceived explanations of length three and four most consistently as concrete, relevant, and coherent. Accordingly, we constructed a new proxy measure CRC for these characteristics by determining whether an explanation fell in that range. At the end of the second iteration, we assessed that the quality of explanations was not yet satisfactory. Still, given the new CRC measure, we were hopeful that another iteration of the Calibration phase would yield a significant improvement.

During the third iteration, we calibrated the XAI system to the new CRC measure while maintaining faithfulness (median length 3 with MAD 0.26, faithfulness 98%). To this end, we adapted the contrast pruning to leave a minimum of three features in any explanation. The generated explanations fell into the target range in 94% of cases, whereas for the previous parametrization, only 25% did. Further, to better convey the magnitude of the contrast, we added more nuanced comparatives based on the difference between fact and foil relative to the features' standard deviation in the dataset. In Step VI, we again conducted a human-grounded evaluation (N=100) through the Clickworker platform. Users judged newly generated explanations against the explanations generated with the previous parametrization of the XAI system as well as human-made explanations as a benchmark. Our analysis in Step VII confirmed the previous finding that concreteness, relevance, and coherence were decisive characteristics for selecting an explanation. In more than two out of three cases (69.1%, $p<0.001$), the explanations generated with the new parametrization were preferred to that of the previous iteration. Users perceived them as more concrete (40.5%, $p<0.001$), more coherent (34.5%, $p<0.05$), and longer (54.3%, $p<0.001$). Overall, we found that the XAI system's explanations outperformed both that of the previous parametrizations as well as the human-made explanations. Thus, we deemed the XAI system fit for deployment.

In the case of a real-world application, in Step IX, the XAI system would be monitored and evaluated throughout its lifecycle. On a technical level, we would continuously monitor the explanations produced by the XAI via a dashboard (Microsoft 2020). If the explanations fell below specified thresholds (e.g., faithfulness below 95%), we would resume the process with Step V. To verify that the XAI system's explanations indeed entertain and educate the app's users as intended, we could occasionally present short surveys to or conduct interviews with randomly selected users of the app.

Evaluation

We evaluate our artifact, the "User-centric XAI process," with respect to its objective and its efficacy (Hevner et al. 2004). As detailed in the previous section, the process succeeded in guiding the instantiation and calibration of a user-centric XAI system that produced explanations that exhibit the identified decisive characteristics. Specifically, the explanations were perceived as concrete, relevant, and coherent by users while being faithful to the explained AI system. We identify the parametrization of the second iteration of the Calibration phase as state of the art (SotA). At this stage, the XAI system's parametrization was informed by recent literature (cf. Instantiation phase) as well re-calibrated (median length 2 with MAD 0.93, generality 98%, faithfulness 100%) based on the results of the user test conducted in the first iteration of the Calibration phase. We argue that this choice of a benchmark is justified due to the absence of previous reports on XAI systems in the application context and the fact that the predominant practice in XAI research to date is to instantiate XAI methods without the involvement of users (Abdul et al. 2018; Wachter et al. 2018). Indeed, the parametrization of the second iteration of the Calibration phase exceeds the current de-

facto standard in XAI research (Abdul et al. 2018; Kirsch 2018; Miller et al. 2017; Mittelstadt et al. 2019). Hence, the second iteration's parametrization of the XAI system represents, if anything, an upper bound on the SotA. Utilizing the data collected in Step VI of the third iteration, we compare the results for our final XAI system with that of the second iteration (SotA) and the human benchmark. More specifically, we analyze the participants' preferences for one out of two different contrastive explanations in a binary choice experiment (Doshi-Velez and Kim 2018). We additionally identify the reasons for participants preferring an explanation, which they select from a pre-defined list of characteristics (cf. Förster et al. 2020). The analysis (cf. Table 2) reveals that the XAI system calibrated with our novel process significantly outperforms the SotA directly as well as in comparison to the human benchmark. As described in detail above, this can be conclusively attributed to the perception of the XAI system's explanations as more concrete (40.5%, $p < 0.001$), more coherent (34.5%, $p < 0.005$), and longer (54.3%, $p < 0.001$).

Table 2. Comparison of State of the Art and Artifact's final XAI system.

	Users prefer explanations over that of competing approach	Users prefer explanations over human benchmark
State of the Art	30.95% ($p < 0.001$)	40.6% ($p < 0.01$)
Artifact's final XAI system	**69.05%** ($p < 0.001$)	**65.3%** ($p < 0.001$)

Results from the user test of iteration 3. p-values given are for a one-sided binomial test (H_0=50%).

Table 2. Comparison of State of the Art and Artifact's final XAI system

To assess the efficacy of the "User-centric XAI process," in the following, we examine each of its three core components (cf. Basic Idea). First, the process proved well-suited to reach the objective. More precisely, we found all phases and steps to be indispensable and placed in a sensible order. In the beginning, the Instantiation phase guided the team from understanding the application context towards the instantiation of a suitable XAI system. The translation of the application requirements into the mode of presentation and desired characteristics provided the essential foundation for the Calibration phase. Entering this phase with a functionally-grounded evaluation ensured that the first parametrization of the XAI system presented to users was already well-tested, in turn enabling the collection of reliable data. Further, the team had the opportunity to familiarize themselves with the potentials and constraints of the XAI system, which added to the first round of analyses. As is evident from the mixed results obtained in the user tests of the first two iterations of the Calibration phase, requiring multiple iterations of calibration and rigorous user testing was invaluable. It ensured that the XAI system verifiably fulfilled all application requirements before it was deemed fit for deployment. Second, incorporating the principles of user-centric design ensured that the user was at the center of attention throughout the process. On a technical level, the XAI system built on a well-known algorithm was capable of generating explanations right after its instantiation. However, as the first user test unambiguously revealed, these explanations failed both to satisfy the users and to meet the requirements. Importantly, the demonstration highlighted that, in line with the quest for fostering human agency, the process incorporates user-centricity beyond user satisfaction. On the one hand, the team emphasized the characteristics users appreciated most when calibrating the XAI system. On the other hand, however, based on the identified needs and expectations of the users, faithfulness was kept as a requirement even when the analyses of user tests revealed that in the test, it was not a decisive consideration for participants. Third, the iterative integration of functionally-grounded and human-grounded evaluation was invaluable. The functionally-grounded evaluation proved indispensable to find a suitable parametrization of the XAI system. We conservatively estimate that across the three iterations of the Calibration phase conducted for the demonstration, we generated and assessed well above 250,000 explanations. On the one hand, it would have been impossible to evaluate even a fraction of these with human users. On the other hand, without validated proxy measures, functionally-grounded evaluation would have been futile. Here, the first iteration of the Calibration phase revealed that despite the thorough assessment of the application context and extensive research in the literature, the first parametrization of the XAI system failed to generate explanations that met the objective. On the contrary, the human-grounded evaluation revealed that the common assumption that users prefer concise explanations did not hold. The second iteration uncovered the relative importance of characteristics and the set of decisive characteristics, enabling us to find an empirically validated proxy measure for concreteness, relevance, and coherence. This proxy measure enabled us to systematically find a parametrization that outperformed the current state of the art both directly and with respect to a common human benchmark (cf. Table 2).

Conclusion, Limitations, and Directions for Further Research

Opacity renders the recommendations of an AI system unintelligible to the user (Guidotti et al. 2019b), which impedes human agency, hinders societal acceptance, and thus poses a critical impediment to exploit AI's potential to benefit individuals and society (HLEG-AI 2019). Indeed, opacity fosters distrust, both reducing users' willingness to accept AI decisions (Herse et al. 2018) as well as inhibiting users' critical reflection before following an AI recommendation (Rader and Gray 2015). In light of this challenge, XAI aims to provide automatically generated explanations for AI systems (HLEG-AI 2019) that enable users to understand, contest, and alter an AI system's decisions (Doran et al. 2018). While a plethora of approaches has been demonstrated in the quest to provide explanations (Adadi and Berrada 2018), XAI research is criticized for not putting the user at the center of attention (Kirsch 2018). While first studies examine insights from the social sciences (Miller 2019) and evaluate explanations from users' perspectives (Förster et al. 2020), a process is needed that effectively guides researchers and practitioners in the design of user-centric XAI systems.

Against this background, we designed the "User-centric XAI process" to systematically guide the instantiation, calibration, and quality control of XAI systems such that they foster human agency and enable appropriate trust in AI systems. Our artifact's sequential structure is based on well-established processes from the fields of data mining and data science (cf. Martinez-Plumed et al. 2019). It incorporates the complementary scenarios for the evaluation of explainable systems proposed by Doshi-Velez and Kim (2018) and the principles of user-centric design (ISO 9241-210 2019), as well as insights from the social sciences into characteristics and presentation modes of explanations appreciated by users (cf. Miller 2019). We demonstrated the practical applicability of our artifact and rigorously evaluated its efficacy in the realistic setting of a smartphone app. Our contribution to research and practice is twofold. First, following the Design Science methodology (Hevner et al. 2004), we conceptualized and evaluated a process that effectively and systematically guides researchers and practitioners in the design of user-centric XAI systems. We contribute to the successful development and application of XAI systems by providing a structure for their design that keeps the user at the center of attention. At the same time, the iterative calibration ensures an appropriate balance between costly user testing and efficient optimization towards well-founded proxy measures. While focusing on post hoc interpretability and model-agnostic XAI methods, our process can be applied independently of the underlying AI system and application domain. Second, we demonstrate how to effectively incorporate processes from data mining and data science, principles of user-centric design, insights from the social sciences, and evaluation frameworks for XAI systems into a unified process. This unification puts research from different disciplines into the context of XAI and the quest for human agency, enabling researchers to identify the broader implications and links between previously fragmented findings.

Although our research provides a substantial step towards the design of user-centric XAI systems that foster human agency, it is subject to several limitations. First, notwithstanding the strength of our experiment, we evaluated our process only for one single use case and did not observe long-term effects. Nevertheless, the artifact is well-founded and does not rely on particular properties of the AI system. We encourage researchers and practitioners to apply and evaluate our process in different domains and especially with different target groups to investigate how the process varies with and can be adapted to suit different levels of the end-users' expertise. In addition, as our use case is built on an application focusing on AI as augmentation, we invite future research to explore the applicability of our process in the context of automated decision-making and interactive AI systems that allow users to contribute their expertise. Overall, studies that observe long-term effects as well as whether the XAI systems designed with our process indeed empower users in real-world applications are of particular interest. Second, translating user requirements into modes of presentation and characteristics and subsequently deriving proxy measures constitute major elements of our process. While a large variety of proxy measures has been reported in the literature, it remains an open question whether proxy measures that truly reflect the users' perception can be identified for all characteristics of explanations. Although our process is well-suited to uncover novel, domain-specific proxy measures, it cannot provide guarantees. Hence, we encourage research both into new proxy measures as well as into the fundamental question whether, in principle, proxy measures can be constructed for any characteristic. Third, designing an XAI system following the "User-centric XAI process" demands significant expenditure of time and costs, even though our artifact is designed mindful of resources, most importantly by aiming for an optimal balance of functionally-grounded and human-

grounded evaluation. As the involvement of users is indispensable for the design of user-centric systems (ISO 9241-210 2019), we invite further research to shed light on alternative means of including users and the associated trade-offs. Finally, given the challenge of human agency and societal acceptance of AI, providing explanations for AI systems and defining processes to tailor them to user requirements constitute an important element but cannot account for all aspects within and beyond the research field of XAI. In particular, the quest for human agency raises challenges for XAI deployments in organizations, which we invite future research to investigate. With our work, we hope to encourage XAI researchers to put the user in the focus of their attention, thereby pushing this fascinating research field forward.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. 2018. "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal, QC.

Adadi, A., and Berrada, M. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access* (6), pp. 52138–52160.

Bandura, A. 2006. "Toward a psychology of human agency," *Perspectives on Psychological Science* (1:2), pp. 164–180.

Baron, J. 2014. "Heuristics and biases," *The Oxford Handbook of Behavioral Economics and the Law,* pp. 3–27.

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., and Herrera, F. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* (58), pp. 82–115.

Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., D'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., and Parekh, J. 2020. "Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach," *SSRN Electronic Journal*.

Buhrmester, M., Kwang, T., and Gosling, S. D. 2011. "Amazon's Mechanical Turk," *Perspectives on Psychological Science* (6:1), pp. 3–5.

Chapman, P., Clinton, J., Kerber, R., Kabaza, T., Reinartz, T., Shearer, C., and Wirth, R. 2000. "CRISP-DM 1.0: Step-by-Step Data Mining Guide," SPSS.

Chen, D. L., Schonger, M., and Wickens, C. 2016. "OTree—An Open-Source Platform for Laboratory, Online, and Field Experiments," *Journal of Behavioral and Experimental Finance* (9), pp. 88–97.

DARPA 2017. "Explainable Artificial Intelligence (XAI)." DARPA. (https://www.darpa.mil/program/explainable-artificial-intelligence, accessed September 4, 2020).

Dhurandhar, A., Pedapati, T., Balakrishnan, A., Chen, P.-Y., Shanmugam, K., and Puri, R. 2019. "Model Agnostic Contrastive Explanations for Structured Data," *ArXiv* (1906.00117).

Doran, D., Schulz, S., and Besold, T. R. 2018. "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives," in *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017*, Bari.

Doshi-Velez, F., and Kim, B. 2018. "Considerations for Evaluation and Generalization in Interpretable Machine Learning," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Cham: Springer, pp. 3–17.

Farinango, C. D., Benavides, J. S., and Lopez, D. M. 2015. "OpenUP/MMU-ISO 9241-210. Process for the Human Centered Development of Software Solutions," *IEEE Latin America Transactions* (13:11), IEEE Computer Society, pp. 3668–3675.

Förster, M., Klier, M., Kluge, K., and Sigler, I. 2020. "Evaluating Explainable Artificial Intelligence – What Users Really Appreciate," in *Proceedings of the European Conference on Information Systems 2020*, Marrakesh.

Gertosio, C., and Dussauchoy, A. 2004. "Knowledge Discovery from Industrial Databases," *Journal of Intelligent Manufacturing* (15:1), pp. 29–37.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*, Cambridge, MA: MIT Press.

Google 2019. "People + AI Guidebook." Google. (https://pair.withgoogle.com/guidebook, accessed September 4, 2020).

Gould, J. D., and Lewis, C. 1985. "Designing for usability: key principles and what designers think," *Communications of the ACM* (28:3), pp. 300–311.

Green, D. P., and Kern, H. L. 2010. "Modeling Heterogeneous Treatment Effects in Large-Scale Experiments Using Bayesian Additive Regression Trees," in *Proc. Annu. Summer Meeting Soc. Political Methodol.*, pp. 1–40.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., and Turini, F. 2019a. "Factual and Counterfactual Explanations for Black Box Decision Making," *IEEE Intelligent Systems* (34:6), pp. 14–23.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. 2019b. "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys* (51:5), pp. 1–42.

Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S., Nottle, A., and Preece, A. 2019. "A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems," in *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, Macau.

Herse, S., Vitale, J., Tonkin, M., Ebrahimian, D., Ojha, S., Johnston, B., Judge, W., and Williams, M. 2018. "Do You Trust Me, Blindly? Factors Influencing Trust Towards a Robot Recommender System," in *27th IEEE International Symposium on Robot and Human Interactive Communication*, Nanjing, pp. 7–14.

Hevner, March, Park, and Ram. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75–105.

Hilton, D. J., and Erb, H.-P. 1996. "Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance," *Thinking & Reasoning* (2:4), pp. 273–308.

HLEG-AI. 2019. "Ethics Guidelines for Trustworthy Artificial Intelligence," Brussels: Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., and Baesens, B. 2011. "An Empirical Evaluation of the Comprehensibility of Decision Table, Tree and Rule Based Predictive Models," *Decision Support Systems* (51:1), pp. 141–154.

IBM. 2016. "Analytics Solutions Unified Method," *Analytics Services Datasheet*, Armonk, NY: IBM Corporation.

IDEO 2015. "The Field Guide to Human-Centred Design." IDEO. (https://www.designkit.org//resources/1, accessed September 4, 2020).

ISO 9241-210. 2019. "Ergonomics of Human-System Interaction — Part 210: Human-Centred Design for Interactive Systems," Geneva: International Organization for Standardization.

Kirsch, A. 2018. "Explain to Whom? Putting the User in the Center of Explainable AI," in *Proceedings of the 1st International Workshop on Comprehensibility and Explanation in AI and ML 2017*, Bari.

Kurgan, L. A., and Musilek, P. 2006. "A Survey of Knowledge Discovery and Data Mining Process Models," *The Knowledge Engineering Review* (21:1), pp. 1–24.

Lighthall, G. K., & Vazquez-Guillamet, C., 2015. "Understanding decision making in critical care," *Clinical Medicine & Research* (13:3-4), pp. 156–168.

Lipton, P. 1990. "Contrastive Explanation," *Royal Institute of Philosophy Supplement* (27), pp. 247–266.

Lipton, P. 2000. "Inference to the Best Explanation," in *A Companion to the Philosophy of Science*, W. H. Newton-Smith (ed.), Maiden, MA: Blackwell, pp. 184–193.

Lipton, Z. C. 2018. "The Mythos of Model Interpretability," *Queue* (16:3), pp. 1–27.

Lombrozo, T. 2012. "Explanation and Abductive Inference," in *The Oxford Handbook of Thinking and Reasoning*, K. J. Holyoak and R. G. Morrisson (eds.), Oxford: Oxford University Press.

Lundberg, S., and Lee, S.-I. 2017. "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, pp. 4765–4774.

Marbán, O., Segovia, J., Menasalvas, E., and Fernández-Baizán, C. 2009. "Toward Data Mining Engineering: A Software Engineering Approach," *Information Systems* (34), pp. 87–107.

Martens, D., and Provost, F. 2014. "Explaining Data-Driven Document Classifications," *MIS Quarterly* (38:1), pp. 73–99.

Martin, K., 2019. "Designing Ethical Algorithms," *MIS Quarterly Executive* (18:2), pp. 129–142.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez Orallo, J., Kull, M., Lachiche, N., Ramirez Quintana, M. J., and Flach, P. A. 2019. "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Transactions on Knowledge and Data Engineering*.

McClure, J. 2002. "Goal-Based Explanations of Actions and Outcomes," *European Review of Social Psychology* (12:1), pp. 201–235.

Microsoft. 2020. "Team Data Science Process Documentation," Microsoft. (https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/, accessed September 4, 2020).

Miller, T. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial*

Intelligence (267), pp. 1–38.
Miller, T., Howe, P., and 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-17 Workshop on Explainable AI (XAI) Proceedings*, Melbourne, pp. 36–42.
Mittelstadt, B., Russell, C., and Wachter, S. 2019. "Explaining Explanations in AI," in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, pp. 279–288.
Mohseni, S., and Ragan, E. D. 2018. "A Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning," *ArXiv* (1801.05075).
Mothilal, R. K., Sharma, A., and Tan, C. 2020. "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, pp. 607–617.
Moyle, S., and Jorge, A. 2001. "RAMSYS - A Methodology for Supporting Rapid Remote Collaborative Data Mining Projects," in *Proceedings of the ECML/PKDD'01 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, Freiburg, pp. 20–31.
Norman, D. A., and Draper, S. W. (eds.). 1986. "User Centered System Design: New Perspectives on Human-Computer Interaction," *User Centered System Design* (1st ed.), Boca Raton, FL: CRC Press.
Nunes, I., and Jannach, D. 2017. "A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems," *User Modeling and User-Adapted Interaction* (27), pp. 393–444.
Preece, A., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. 2018. "Stakeholders in Explainable AI," in *Artificial Intelligence in Government and Public Sector Proceedings*, Arlington, VA.
Rader, E., and Gray, R. 2015. "Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul, pp. 173–182.
Ramcharan, A., McCloskey, P., Baranowski, K., Mbilinyi, N., Mrisho, L., Ndalahwa, M., Legg, J., and Hughes, D. P. 2019. "A Mobile-Based Deep Learning Model for Cassava Disease Diagnosis," *Frontiers in Plant Science* (10), pp. 1-8.
Rai, A. 2020. "Explainable AI: From black box to glass box," *Journal of the Academy of Marketing Science* 48(1), pp. 137-141.
Ribera, M., and Lapedriza, A. 2019. "Can We Do Better Explanations? A Proposal of User-Centered Explainable AI," in *Joint Proceedings of the ACM IUI 2019 Workshops*, Los Angeles, CA.
Rollins, J. B. 2015. "Foundational Methodology for Data Science," *IBM Analytics Whitepaper*, Somers, NY: IBM Corporation.
Silva, P. F. B., Marcal, A. R. S., and da Silva, R. A. 2014. "Leaf Dataset," *UCI Machine Learning Repository*. (https://archive.ics.uci.edu/ml/datasets/Leaf, accessed April 30, 2020).
Silva, P. F. B., Marçal, A. R. S., and da Silva, R. M. A. 2013. "Evaluation of Features for Leaf Discrimination," in *International Conference Image Analysis and Recognition*, Póvoa do Varzim, pp. 197–204.
Sonnenberg, C., and vom Brocke, J. 2012. "Evaluations in the Science of the Artificial – Reconsidering the Build-Evaluate Pattern in Design Science Research," in *Design Science Research in Information Systems. Advances in Theory and Practice*, Las Vegas, NV, pp. 381–397.
Still, B., and Crane, K. 2017. *Fundamentals of User-Centered Design: A Practical Approach*, Boca Raton, FL: CRC Press.
Thagard, P. 1989. "Explanatory Coherence," *Behavioral and Brain Sciences* (12), pp. 435–467.
Tversky, A., and Kahneman, D. 2015. "Causal schemas in judgments under uncertainty," in *Progress in Social Psychology: Volume 1*, New York, NY: Psychology Press, pp. 49–72.
Wachter, S., Mittelstadt, B., and Russell, C. 2018. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology* (31:2), pp. 841–887.
Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. 2019. "Designing Theory-Driven User-Centric Explainable AI," in *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, pp. 1–15.
Weerts, H. J. P., van Ipenburg, W., and Pechenizkiy, M. 2019. "A Human-Grounded Evaluation of SHAP for Alert Processing," in *Proceedings of the KDD Workshop on Explainable AI*, Anchorage, AK.
Wolf, C. T. 2019. "Explainability Scenarios: Towards Scenario-Based XAI Design," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Marina del Rey, CA, pp. 252–257.

3.3 Explainable Artificial Intelligence in Information Systems – A Review of the Status Quo and Future Research Directions

Full Citation:	Brasse, J., Broder, H.R., Förster, M. et al. Explainable artificial intelligence in information systems: A review of the status quo and future research directions. Electron Markets 33, 26 (2023). https://doi.org/10.1007/s12525-023-00644-5
Copyright Note:	

Electronic Markets (2023) 33:26
https://doi.org/10.1007/s12525-023-00644-5

RESEARCH PAPER

Explainable artificial intelligence in information systems: A review of the status quo and future research directions

Julia Brasse[1] · Hanna Rebecca Broder[1] · Maximilian Förster[1] · Mathias Klier[1] · Irina Sigler[1]

Received: 31 July 2022 / Accepted: 30 March 2023

Abstract

The quest to open black box artificial intelligence (AI) systems evolved into an emerging phenomenon of global interest for academia, business, and society and brought about the rise of the research field of explainable artificial intelligence (XAI). With its pluralistic view, information systems (IS) research is predestined to contribute to this emerging field; thus, it is not surprising that the number of publications on XAI has been rising significantly in IS research. This paper aims to provide a comprehensive overview of XAI research in IS in general and electronic markets in particular using a structured literature review. Based on a literature search resulting in 180 research papers, this work provides an overview of the most receptive outlets, the development of the academic discussion, and the most relevant underlying concepts and methodologies. Furthermore, eight research areas with varying maturity in electronic markets are carved out. Finally, directions for a research agenda of XAI in IS are presented.

Keywords Explainable artificial intelligence · Explainable machine learning · Comprehensible artificial intelligence · Comprehensible machine learning · Literature review

JEL Classification M10

Introduction

Artificial intelligence (AI) is already ubiquitous at work and in everyday life: in the form of diverse technologies, such as natural language processing or image recognition (Abdul et al., 2018; Berente et al., 2021) and in various application domains, including electronic markets, finance, healthcare, human resources, public administration, and transport (Collins et al., 2021; Meske et al., 2020). The presence of AI will expand as about 70% of companies worldwide intend to adopt AI by 2030 (Bughin et al., 2018). Thereby, AI is expected to transform all aspects of society (Collins et al., 2021; Makridakis, 2017).

The current CEO of Alphabet Inc. anticipates AI to "have a more profound impact on humanity than fire, electricity and the internet" (Knowles, 2021). AI holds great potential through tremendous efficiency gains and novel information processing capabilities (Asatiani et al., 2021) and even surpasses human performance in specific tasks (Meske et al., 2022). For instance, AI has outperformed physicians in diagnosing breast cancer (e.g., McKinney et al., 2020). At the same time, the use of AI is associated with severe risks, particularly concerning managerial issues such as inscrutability, ethical issues including fairness, justice, and discrimination, and legal issues such as accountability, regulation, and responsibility (Akter et al., 2021a; Asatiani et al., 2021; Berente et al., 2021). Potential negative consequences of AI usage affect not only individuals and organizations, but society as a whole (Mirbabaie et al., 2022; Robert et al., 2020). For example,

Responsible Editor: Shahriar Akter

✉ Mathias Klier
mathias.klier@uni-ulm.de

Julia Brasse
julia.brasse@uni-ulm.de

Hanna Rebecca Broder
hanna.broder@uni-ulm.de

Maximilian Förster
maximilian.foerster@uni-ulm.de

Irina Sigler
irina.sigler@uni-ulm.de

[1] Institute of Business Analytics, University of Ulm, Helmholtzstraße 22, 89081 Ulm, Germany

Published online: 27 May 2023

an AI-based debt recovery program called "Robodebt" scheme unlawfully claimed almost $2 billion from more than 400,000 Australian citizens (Australian Broadcasting Corporation, 2022). There are growing concerns that using AI could exacerbate social or economic inequalities (Gianfrancesco et al., 2018). Examples include an AI-based recruiting engine used by Amazon.com Inc. which downgraded resumes from female in favor of male candidates (Gonzalez, 2018), an AI operated by Twitter Inc. to communicate with users who became verbally abusive, and an AI used by Google LLC which returned racist results in image searches (Yampolskiy, 2019).

The advancing capabilities of AI models contribute to their opacity, rendering their functioning and results uninterpretable to humans (Berente et al., 2021). Opacity can, one the one hand, lead to humans blindly relying on AI results and substituting their own judgment with potentially false decisions (Robert et al., 2020). On the other hand, the lack of interpretability may lead to reluctance to use AI. In the case of breast cancer diagnosis, AI-based decision support systems may fail to detect certain diseases, for instance, due to biased training data. Physicians exhibiting overreliance may fail to detect these errors; physicians that do not trust AI systems and refuse to use them may not benefit from the decision support.

Explainable AI (XAI) aims at both leveraging the potential and mitigating the risks of AI by increasing its explainability. XAI aims to empower human stakeholders to comprehend, appropriately trust, and effectively manage AI (Arrieta et al., 2020; Langer et al., 2021). In the example of breast cancer diagnosis, explainability can assist physicians in understanding the functioning and results of an AI-based decision support system. Thus, it may help them appropriately trust the system's decisions and detect its errors. Consequently, a partnership between physicians and AI might make better decisions than either physicians or AI individually. Efforts to increase the explainability of AI systems are emerging across various sectors of society. Companies strive to make their AI systems more comprehensible (e.g., Google, 2022; IBM, 2022). Regulators take action to demand accountability and transparency of AI-based decision processes. For instance, the European General Data Protection Regulation (GDPR) guarantees the "right to explanation" for those affected by algorithmic decisions (Selbst & Powles, 2017). The upcoming EU AI regulation requires human oversight—to interpret and contest AI systems' outcomes—in "high-risk" applications such as recruiting or creditworthiness evaluation (European Commission, 2021). XAI's economic and societal relevance attracts researchers' attention, which manifests in an increasing number of publications in recent years (Arrieta et al., 2020). For instance, XAI researchers work on revealing the functioning of specific AI-based applications, such as cancer diagnosis systems (Kumar et al., 2021) and malware prediction systems (Iadarola et al., 2021), to their users. Further, they investigate approaches to automatically generate explanations along AI decisions that can be applied independently from the underlying AI model. Exemplary use cases include credit risk assessment (Bastos & Matos, 2021) or fraud detection (Hardt et al., 2021). Information systems (IS) research is predestined to investigate and design AI explainability, as it views technology from individuals', organizations', and society's perspectives (Bauer et al., 2021).

Especially for an emerging research field such as XAI, a literature review can help to create "a firm foundation for advancing knowledge" (Webster & Watson, 2002, p. 13) and put forward the research's relevance and rigor (vom Brocke et al., 2009). We aim to provide deeper insights into this body of knowledge by conducting a structured literature review. The contribution is twofold: First, we provide a structured and comprehensive literature review of XAI research in IS. Second, we provide a future research agenda for XAI research in IS.

Our paper is structured as follows: In the following, we provide an overview of related work and outline our research questions. In the third section, we present the methodology, followed by the results in the fourth section. Finally, we carve out a future research agenda and present the contribution, implications, and limitations.

Theoretical background and related work

Theoretical foundations

Given that IS research investigates and shapes "how individuals, groups, organizations, and markets interact with IT" (Sidorova et al., 2008, p. 475), human-AI interaction is a crucial research topic for the discipline. In general, human-agent interaction occurs between an IT system and a user seeking to conduct a specific task in a given context (Rzepka & Berger, 2018). It is determined by the characteristics of the task, the context, the user, and the IT system (Rzepka & Berger, 2018). When the human counterpart is an AI system, specific characteristics of AI systems must be considered. Modern AI systems with continually evolving frontiers of emerging computing capabilities provide greater autonomy, more profound learning capacity, and higher inscrutability than previously studied IT systems (Baird & Maruping, 2021; Jiang et al., 2022). The rapid progress in AI is primarily contributed to the rise of machine learning (ML), which can be defined as the ability to learn specific tasks by constructing models based on processing data (Russell & Norvig, 2021). The autonomy and learning capacity of ML-based AI systems further reinforce inscrutability (Berente et al., 2021). Thus,

challenges arise to manage human-AI interaction with ever-increasing levels of AI autonomy, learning capacity, and inscrutability.

From a managerial perspective, inscrutability carries four interdependent emphases: opacity, transparency, explainability, and interpretability (Berente et al., 2021). First, opacity is a property of the AI system and refers to its complex nature, which impedes humans from understanding AI's underlying reasoning processes (Meske et al., 2020). Many AI systems are "black boxes," which means that the reasons for their outcomes remain obscure to humans—often not only to the users but also to the developers (Guidotti et al., 2019; Merry et al., 2021). A prominent example are neural networks. Second, transparency refers to the willingness to disclose (parts of) the AI system by the owners and is thus considered a strategic management issue (Granados et al., 2010). Third, explainability is a property of the AI system and refers to the system's ability to be understood by at least some parties, at least to a certain extent (Gregor & Benbasat, 1999). Finally, interpretability refers to the understandability of an AI system from human perspectives. An AI system with a certain degree of explainability might be adequately interpretable for one person but not necessarily for another (Berente et al., 2021). For instance, decision trees can become uninterpretable for some users as complexity increases (Mittelstadt et al., 2019).

Opacity significantly affects human-AI interaction: It prevents humans from scrutinizing or learning from an AI system's decision-making process (Arrieta et al., 2020). Confronted with an opaque system, humans cannot build appropriate trust; they often either blindly follow the system's decisions and recommendations or do not use the system (Herse et al., 2018; Rader & Gray, 2015). Thus, opacity constitutes an impediment to both human agency and AI adoption. The research field of XAI addresses the opacity of AI systems. XAI aims at approaches that make AI systems more explainable—sometimes also referred to as comprehensible (Doran et al., 2018)—by automatically generating explanations for their functioning and outcomes while maintaining the AI's high performance levels (Adadi & Berrada, 2018; Gregor & Benbasat, 1999). In day-to-day human interaction, "explanation is a social and iterative process between an explainer and an explainee" (Chromik & Butz, 2021, p. 1). This translates into the context of human-AI interaction, where explanations constitute human-understandable lines of reasoning for why an AI system connects a given input to a specific output (Abdul et al., 2018). Thus, explanations can address the opacity of AI systems and increase their interpretability from users' perspectives. Researchers emphasize that clarifying XAI's role can make significant contributions to the ongoing discussion of human-AI interaction (Sundar, 2020).

Terminological foundations

The XAI research discipline is driven by four key goals (Adadi & Berrada, 2018; Arrieta et al., 2020; Gerlings et al., 2021; Gilpin et al., 2018; Langer et al., 2021; Meske et al., 2020; Wang et al., 2019): First, to generate explanations that allow to *evaluate* an AI system and thus detect its flaws and prevent unwanted behavior (Adadi & Berrada, 2018; Gerlings et al., 2021; Meske et al., 2020; Wang et al., 2019). For instance, evaluation in this context is utilized to detect and prevent non-equitable treatment of marginalized communities (Arrieta et al., 2020). The second goal is to build explanations that help to *improve* an AI system. In this case, explanations can be used by developers to improve a model's accuracy by deepening their understanding of the AI system's functioning (Adadi & Berrada, 2018; Arrieta et al., 2020; Gilpin et al., 2018; Langer et al., 2021; Meske et al., 2020). Third, to provide explanations that *justify* an AI system's decisions by improving transparency and accountability (Adadi & Berrada, 2018; Gerlings et al., 2021; Meske et al., 2020; Wang et al., 2019). One prominent example highlighting the need to justify is based on the "right to explanation" for those affected by algorithmic decisions (cf., e.g., GDPR); another example concerns decisions made by a professional who follows an AI system's recommendation but remains accountable for the decision (Arrieta et al., 2020). Finally, to produce explanations that allow to *learn* from the system by unmasking unknown correlations that could indicate causal relationships in the underlying data (Adadi & Berrada, 2018; Langer et al., 2021; Meske et al., 2020). In a nutshell, XAI aims to evaluate, improve, justify, and learn from AI systems by building explanations for a system's functioning or its predictions (Abdul et al., 2018; DARPA, 2018).

To reach these goals, XAI research provides a wide array of approaches that can be grouped along two dimensions: scope of explainability and model dependency (Adadi & Berrada, 2018; Arrieta et al., 2020; Vilone & Longo, 2020). The scope of explainability can be global or local (Adadi & Berrada, 2018; Arrieta et al., 2020; Heuillet et al., 2021; Payrovnaziri et al., 2020; Vilone & Longo, 2020). A *global* explanation targets the functioning of the entire AI model. Using the example of credit line decisions, a global explanation might highlight the most relevant criteria that are exploited by the AI model to derive credit line decisions. *Local* explanations, on the other hand, focus on rationalizing an AI model's specific outcome. Returning to the example of credit line decisions, a local explanation might provide the most essential criteria for an individual denial or approval. The second dimension, dependency on the AI model, distinguishes between two approaches: model-specific and model-agnostic (Adadi & Berrada, 2018; Arrieta et al., 2020; Rawal et al., 2021). *Model-specific* approaches focus on providing

explanations for specific AI models or model classes (Arrieta et al., 2020; Rawal et al., 2021), like neural networks (Montavon et al., 2018), as they consider internal components of the AI model (class), such as structural information. In turn, *model-agnostic* approaches disregard the models' internal components and are thus applicable across a wide range of AI models (Adadi & Berrada, 2018; Rawal et al., 2021; Ribeiro et al., 2016; Vilone & Longo, 2020).

Designing or choosing the best XAI approach for a given problem is equivalent to solving a "human-agent interaction problem" (Miller, 2019, p. 5). Thus, it is vital to consider an explanation's audience. Three major target groups are the focus of XAI research (Bertrand et al., 2022; Cooper, 2004; Ribera & Lapedriza, 2019; Wang et al., 2019). The first group comprises *developers* who build AI systems, i.e., data scientists, computer engineers, and researchers (Bertrand et al., 2022; Ribera & Lapedriza, 2019; Wang et al., 2019). To illustrate, using the example of credit line decisions, this is the team building the AI system or responsible for maintaining it. The second group contains *domain experts* who share expertise based on formal education or professional experience in the application field (Bertrand et al., 2022; Ribera & Lapedriza, 2019; Wang et al., 2019). In the case of credit line decisions, this would be the bank advisor accountable for the credit line decision. The final group, *lay users*, includes individuals who are affected by AI decisions (Bertrand et al., 2022; Cooper, 2004; Ribera & Lapedriza, 2019), e.g., the bank customer who was approved or denied a credit line based on an AI system's recommendation (Mittelstadt et al., 2019). Additionally, this third group includes lay users that interact with an AI, e.g., customers who explore credit line options with the help of an AI-based agent.

To investigate to what extent XAI approaches solve this "human-agent interaction problem," literature established a baseline of three different evaluation scenarios (Adadi & Berrada, 2018; Chromik & Schuessler, 2020; Doshi-Velez & Kim, 2018). *Functionally grounded evaluation*, as the first scenario, is employed to assess the technical feasibility of XAI approaches and explanations' characteristics employing proxy measures (Doshi-Velez & Kim, 2018), e.g., analyze an explanation's length to assess its complexity (Martens & Provost, 2014; Wachter et al., 2018). While functionally grounded evaluation omits user involvement, both the second and the third scenarios build on studies with humans (Doshi-Velez & Kim, 2018). The second scenario, *human-grounded evaluation*, aims to assess the quality of explanations by conducting studies with human subjects who are not necessarily the target users, e.g., students, performing simplified proxy-tasks (Doshi-Velez & Kim, 2018; Förster et al., 2020a). *Application-grounded evaluation*, as the third scenario, is based on real-world testing involving the intended users of an AI system and deployment in the actual application setting (Abdul et al., 2018). Reverting to the example of the credit line decisions, an application-grounded evaluation would be set in an actual bank environment, with actual bank advisors and/or customers as subjects, while human-grounded evaluation would allow for a simulated environment. Table 1 provides an overview of key concepts and definitions in XAI research, which we will draw on when analyzing the identified body of literature for providing a comprehensive literature review of XAI research in IS.

Existing literature reviews on XAI

Several literature reviews address the growing body of research in the field of XAI applying different foci and angles. While some of them aim at formalizing XAI (e.g., Adadi & Berrada, 2018), for example, by drawing together the body of knowledge on the nature and use of explanations from intelligent systems (Gregor & Benbasat, 1999), others provide taxonomies for XAI in decision support (Nunes and Jannach, 2017) or survey methods for explaining AI (e.g., Guidotti et al., 2019). Other literature reviews focus on specific (X)AI methods, such as rule-based models (e.g., Kliegr et al., 2021), neuro-fuzzy rule generation algorithms (e.g., Mitra & Hayashi, 2000), or neural networks (e.g., Heuillet et al., 2021), or review-specific explanation formats, like visual explanations (e.g., Zhang & Zhu, 2018). Another stream of literature reviews highlights user needs in XAI, for example, by reviewing design principles for user-friendly explanations (Chromik & Butz, 2021) or XAI user experience approaches (Ferreira & Monteiro, 2020).

Another group of literature reviews on XAI focuses on specific application domains like healthcare (e.g., Amann et al., 2020; Chakrobartty & El-Gayar, 2021; Payrovnaziri et al., 2020; Tjoa & Guan, 2021), finance (e.g., Kute et al., 2021; Moscato et al., 2021), or transportation (e.g., Omeiza et al., 2021). For example, Amann et al. (2020) provide a comprehensive review of the role of AI explainability in clinical practice to derive an evaluation of what explainability means for the adoption of AI-based tools in medicine. Omeiza et al. (2021) survey XAI methods in autonomous driving and provide a conceptual framework for autonomous vehicle explainability. Other scholars apply XAI to adjacent disciplines (e.g., Abdul et al., 2018; Miller, 2019). For instance, in an often-cited paper, Miller (2019) argues that XAI research can build on insights from the social sciences. The author reviews papers from philosophy and psychology which study how people define, generate, select, evaluate, and present explanations and which cognitive biases and social norms play a role. Thereby, most literature reviews describe existing research gaps and point toward future research directions focusing on their specific view.

As outlined above, existing literature reviews cover various aspects of XAI research. However, to our best knowledge, none of them has provided a comprehensive literature review on XAI research in IS. Our literature review aims at addressing this gap.

Table 1 Key concepts in XAI research

Concept	Definition	Source
Dependency on the AI model		
Model-specific	Approaches that focus on providing explanations for specific AI models or model classes	Adadi & Berrada, 2018; Arrieta et al., 2020; Rawal et al., 2021
Model-agnostic	Approaches that disregard the underlying AI model's internal components and are thus applicable across a wide range of AI models	Adadi & Berrada, 2018; Rawal et al., 2021; Ribeiro et al., 2016; Vilone & Longo, 2020
Scope of explainability		
Global explainability	An explanation that targets explaining the functioning of the entire AI model	Adadi & Berrada, 2018; Arrieta et al., 2020; Heuillet et al., 2021; Payrovnaziri et al., 2020; Vilone & Longo, 2020
Local explainability	An explanation that focuses on rationalizing a specific outcome of an AI model	Adadi & Berrada, 2018; Arrieta et al., 2020; Heuillet et al., 2021; Payrovnaziri et al., 2020; Vilone & Longo, 2020
Explanation's target group		
Developers and AI researchers	Data scientists, computer engineers, and researchers who build or maintain AI systems	Bertrand et al., 2022; Ribera & Lapedriza, 2019; Wang et al., 2019
Domain experts	Experts who share expertise in the field of application based on formal education or professional experience	Bertrand et al., 2022; Ribera & Lapedriza, 2019; Wang et al., 2019
Lay users	Non-expert individuals who are affected by AI decisions or who interact with AI systems	Bertrand et al., 2022; Cooper, 2004; Ribera & Lapedriza, 2019
Explanation's goal		
Evaluate the system	Evaluate an AI system to detect its flaws and prevent unwanted behavior	Arrieta et al., 2020; Adadi & Berrada, 2018; Gerlings et al., 2021; Meske et al., 2020; Wang et al., 2019
Improve the system	Improve an AI system's accuracy by deepening the understanding of the AI system's functioning	Adadi & Berrada, 2018; Arrieta et al., 2020; Gilpin et al., 2018; Langer et al., 2021; Meske et al., 2020
Justify the system	Justify an AI system's decisions by improving transparency and accountability	Adadi & Berrada, 2018; Arrieta et al., 2020; Gerlings et al., 2021; Meske et al., 2020; Wang et al., 2019
Learn from the system	Learn from the AI system by identifying unknown correlations that could indicate causal relationships in the underlying data	Adadi & Berrada, 2018; Langer et al., 2021: Meske et al., 2020

Research questions

While considerable progress in XAI has already been made by computer scientists (Arrieta et al., 2020), interest in this field has increased rapidly among IS scholars in recent years (Meske et al., 2020). This is underpinned, for instance, by an increasing number of Calls for Papers (cf., e.g., Special Issue on Explainable and Responsible Artificial Intelligence in *Electronic Markets*, Special Issue on Designing and Managing Human-AI Interactions in *Information Systems Frontiers*), conference tracks (cf., e.g., Minitrack on Explainable Artificial Intelligence at *Hawaii International Conference on System Sciences*), and Editorials (cf., e.g., Editorial "Expl(AI)n It to Me – Explainable AI and Information Systems Research" in *Business & Information Systems Engineering*). In their Editorial, Bauer et al. (2021) emphasize that IS research is predestined to focus on XAI given the versatility of requirements and consequences of explainability from individuals' and society's perspectives. Moreover, in a research note summarizing existing IS journal articles, Meske et al. (2020) call for a resurgence of research on explainability in IS—after explanations for relatively transparent expert systems have been intensively investigated. To the best of our knowledge, no work exists synthesizing XAI research in IS based on a structured and comprehensive literature search.

To provide deeper insights into the research field of XAI in the IS community, we conduct a structured and comprehensive literature review. Our literature review addresses the following research questions (RQ):

RQ1: How can the academic discussion on XAI in the IS literature be characterized?
RQ2: Which are potential future XAI research areas in IS?

To address the first research question, we aim to (i) identify IS publication outlets that are receptive to XAI research, (ii) describe how the academic discussion on XAI in the IS literature developed over time, (iii) analyze the underlying concepts and methodological orientations of the academic discussion on XAI in the IS literature, and (iv) present the most critical XAI research areas in IS literature. To address the second research question, we aim to derive directions for a research agenda of XAI in IS.

Literature review approach

Relying on the previous discussions, we investigate how IS scholars conduct XAI research. We aim at not only summarizing but analyzing and critically examining the status quo of XAI research in IS (Rowe, 2014). This analysis requires a systematic and structured literature review (Bandara et al., 2011; Webster & Watson, 2002). In preparation, it is necessary to apply a comprehensive and replicable literature search strategy, which includes relevant journals and conferences, appropriate keywords, and an adequate time frame (vom Brocke et al., 2009). Bandara et al. (2011) propose two main steps: selecting the relevant sources to be searched (cf. Webster & Watson, 2002) and defining the search strategy in terms of time frame, search terms, and search fields (Cooper, 1988; Levy & Ellis, 2006). In order to systematically analyze the papers according to XAI theory and IS methodology, we added a third step and coded the articles with respect to relevant concepts in the literature (Beese et al., 2019; Jiang & Cameron, 2020).

Source selection

The literature search needs to include the field's leading journals known for their high quality and will thus publish the most relevant research contributions (Webster & Watson, 2002). The renowned Association for Information Systems (AIS), with members from approximately 100 countries, publishes the Senior Scholars' Basket of Journals, as well as the Special Interest Groups (SIG) Recommended Journals. In our search, we included the eight journals in the AIS Senior Scholars' Basket of Journals, and the 64 AIS SIG Recommended Journals. Because of their high quality, we considered all remaining journals in the AIS eLibrary (including Affiliated and Chapter Journals). In order to identify high-quality journals, different rankings are helpful (Akter et al., 2021b; Levy & Ellis, 2006; vom Brocke et al., 2009). We explicitly considered journals from three prominent rankings: First, journals from the Chartered Association of Business Schools (ABS)/Academic Journal Guide (AJG) 2021 (ranking tier 3/4/4* benchmark, category "Information Management"). Second, journals from the Journal Quality List of the Australian Business Deans Council (ABDC) (ranking tier A/A* benchmark, category "Information Systems"). Third, journals from the German Academic Association of Business Research VHB-JOURQUAL3 (ranking tier A + /A/B benchmark, category "Information Systems").

Moreover, it is recommended to include high-quality conference proceedings (Webster & Watson, 2002), especially when analyzing a relatively nascent and emerging research field such as XAI. Conferences are a venue for idea generation and support the development of new research agendas (Levy & Ellis, 2006; Probst et al., 2013). Thus, we included the major international IS conferences. More precisely, we considered the proceedings of the four AIS Conferences and the proceedings of the twelve AIS Affiliated Conferences. In addition, we ensured that all conferences from the VHB-JOURQUAL3 (ranking tier A + /A/B benchmark, category "Information Systems") are included.

This resulted in 105 journals and 17 conferences as sources for our search.

Search strategy and results

The development of XAI as a research field started in the 1970s and gained momentum in the past 5 to 10 years (Adadi & Berrada, 2018; Mueller et al., 2019). In order to gain an overview of the development of XAI research in IS, we chose to not limit the literature search's time frame. To identify relevant publications, we conducted a search using different terms describing XAI via databases that contain the journals and conferences discussed above. Based on terms that are used synonymously to describe research in the field of XAI (cf. Section "Theoretical background and related work"), we determined the following search string to cover relevant articles: ("explainable" AND "artificial intelligence") OR ("explainable" AND "machine learning") OR ("comprehensible" AND "artificial intelligence") OR ("comprehensible" AND "machine learning"). We searched for these terms in the title, abstract, and keywords. Where a search in title, abstract, and keywords was impossible, we applied a full-text search. Please see Fig. 1 for an overview of our search and screening process.

Our literature search, which was performed in January 2022, resulted in 1724 papers. Papers were screened based on titles and abstracts, with researchers reading the full text where necessary. We excluded all papers that did not deal with XAI as defined above. More specifically, we excluded all papers that focus entirely on AI without the notion of explanations. For instance, we excluded papers on how humans can explain AI for other humans. Further, we excluded papers focusing on the explainability of "Good Old Fashioned AI" such as expert or rule-based systems (Meske et al., 2020, p. 6). In contrast to our understanding of AI, as defined in the introduction, this broader definition of AI also includes inherently interpretable systems, such as knowledge-based or expert systems, which do not face the same challenges of lacking transparency.

To determine our data set of relevant papers, three researchers coded independently from each other and discussed coding disagreements to reach consent. At least two researchers analyzed each paper. Interrater reliability measured by Cohen's Kappa was 0.82—"almost perfect agreement" (Landis & Koch, 1977, p. 165). This procedure led to a set of 154 papers, which then served as the basis for a backward (resulting in 32 papers) and forward search (resulting in 28 papers), as suggested by Webster and Watson (2002).

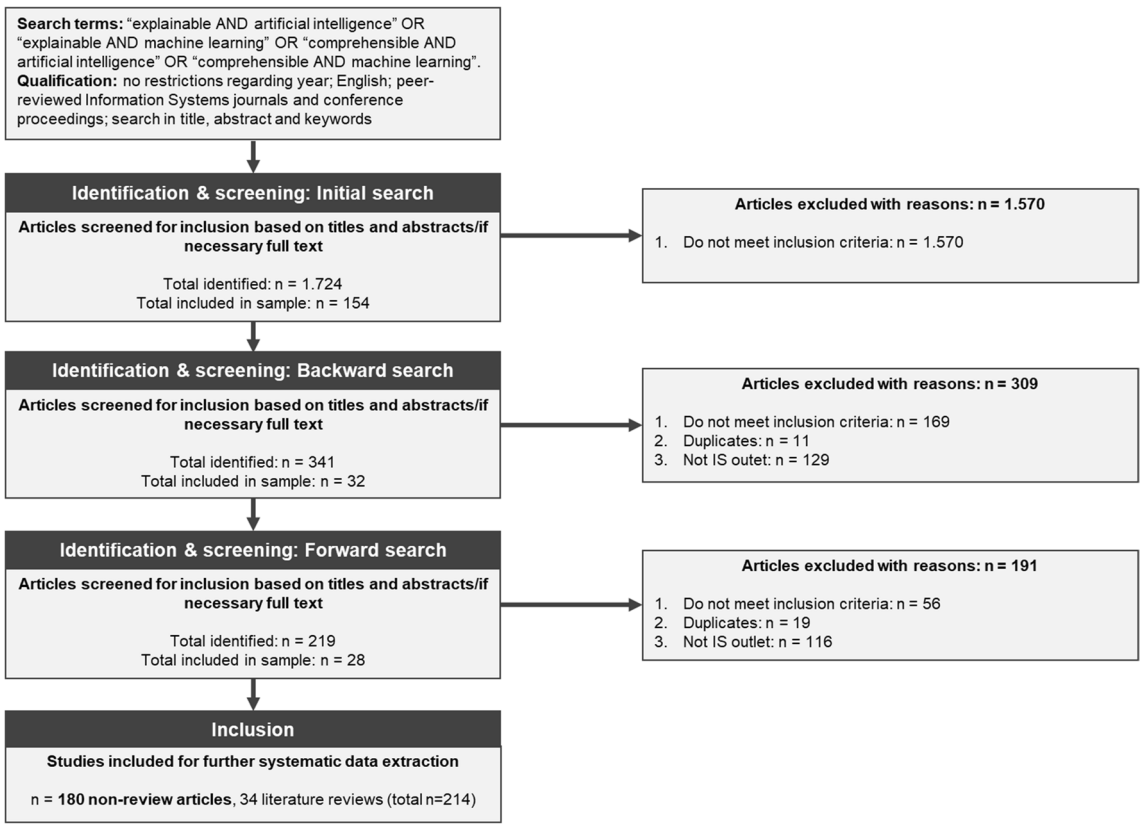


Fig. 1 Search strategy and screening process

We reached a final set of 214 papers that served as the basis for our subsequent analyses.

Analysis scheme and coding procedure

Our goal is to not only summarize but analyze and critically examine the status quo of XAI research in IS (Beese et al., 2019; Rowe, 2014). In order to do so, we first analyzed all 34 papers that solely provide an overview of current knowledge, i.e., literature reviews. We then coded the 180 remaining articles using an analysis scheme derived from existing literature (cf. Section "Terminological foundations"). More specifically, in our analysis, we differentiate relevant theoretical concepts in XAI research and central methodological concepts of IS research. Regarding relevant concepts of XAI literature, we distinguish an XAI approach's dependency on the AI model (Adadi & Berrada, 2018; Arrieta et al., 2020) and its scope of explainability (Adadi & Berrada, 2018; Arrieta et al., 2020; Payrovnaziri et al., 2020; Vilone & Longo, 2020) as well as explanation's target group (Ribera & Lapedriza, 2019; Wang et al., 2019) and goal (Meske et al., 2020). Regarding IS methodology, we distinguish the prevalent research paradigms, i.e., Design Science and Behavioral Science (Hevner et al., 2004). For Design Science contributions, we further specify the artifact type according to Hevner et al. (2004) and the evaluation type according to established evaluation scenarios for XAI approaches (Adadi & Berrada, 2018; Chromik & Schuessler, 2020; Doshi-Velez & Kim, 2018). This results in the following analysis scheme (Fig. 2):

Three researchers coded the 180 remaining articles according to the analysis scheme. Multiple labels per dimension were possible. For a subset of 100 articles, each article was coded by at least two researchers. Interrater reliability measured by Cohen's Kappa was 0.74, which is associated with "substantial agreement" (Landis & Koch, 1977, p. 165). In case of disagreement, the researchers reached a consensus based on discussion.

Results

This section is dedicated to our results. First, we analyze receptive IS publication outlets to XAI research. Second, we examine the development of the academic discussion on XAI in IS literature over time. Third, we analyze the academic discussion's underlying concepts and methodological orientation. Finally, we derive major XAI research areas.

Receptive IS outlets to XAI research

We analyzed which journals and conferences are receptive to XAI research. The results are helpful in three ways: they provide researchers and practitioners with potential outlets

Category	XAI conceptual dimensions			
Dependency on the AI model Adadi and Berrada 2018; Arrieta et al. 2019	**1.** Model-agnostic	**2.** Model-specific		
Scope of explainability Adadi and Berrada 2018; Arrieta et al. 2019; Payrovnaziri et al. 2020; Vilone and Longo 2020	**1.** Local explainability	**2.** Global explainability		
Explanation's target group Ribera and Lapedriza 2019; Wang et al. 2019	**1.** Developers	**2.** Domain experts	**3.** Lay users	
Explanation's goal Adadi and Berrada 2018; Meske et al. 2020	**1.** Evaluate the system	**2.** Improve the system	**3.** Justify the system	**4.** Learn from the system

Category	IS methodological dimensions			
Research paradigm Hevner et al. 2004	**1.** Behavioral Science	**2.** Design Science		
Artifact type Hevner et al. 2004	**1.** Construct	**2.** Model	**3.** Method	**4.** Instantiation
Evaluation type Adadi and Berrada 2018; Chromik and Schussler 2020; Doshi-Velez and Kim 2018	**1.** Functionally-grounded evaluation	**2.** Human-grounded evaluation	**3.** Application-grounded evaluation	

Fig. 2 Analysis scheme

where they can find related research, they assist researchers in identifying target outlets, and they offer insights for editors to what extent their outlet is actively involved in the academic discussion on the topic (Bandara et al., 2011). One hundred forty-one articles were published in journals, and 39 articles in conference proceedings. An overview of the number of publications per journal and per conference is included in the Appendix.

Development of the academic discussion on XAI in IS literature over time

To examine the development of the academic discussion on XAI in IS literature over time, we evaluated the number of articles in conferences and journals per year (cf. Fig. 3). The amount of research increased over time, with the number of publications rising to 79 articles in 2021. Especially from 2019 onward, the number of published articles increased rapidly, with 79% of the studies appearing between 2019 and 2021. The rapid increase since 2019 is not attributed to particular calls for papers or individual conferences but due to a widely growing interest in XAI. In sum, the number of publications per year indicates that the nascent research field of XAI has been gaining significant attention from IS scholars in the last 3 years.

Characteristics of the academic discussion on XAI in IS literature

To examine the characteristics of the academic discussion on XAI in IS literature, we analyzed the dimensions of the research papers according to our analysis scheme, i.e., underlying XAI concepts and methodological orientation (cf. Fig. 4). Note that multiple answers or no answers per category were possible.

Most papers conceptually focus on XAI methods that generate explanations for specific AI systems, i.e., model-specific XAI methods (53%). In contrast, fewer papers deal with model-agnostic XAI methods, which can be used independently of the specific AI system (38%). The scope of explainability under investigation varies: Local explanations that focus on rationalizing an AI system's specific outcome are represented almost equally (55%) to global explanations that examine the functioning of the underlying AI model (57%). Thirty-three articles (18%) feature a combination of local and global explanations. First and foremost, explanations address domain experts (62%), followed by lay users (33%). The predominant goal of XAI is to justify an AI system's decisions (83%).

Regarding methodological orientation, IS research efforts concentrate on developing novel XAI artifacts (76%). Researchers mainly rely on the functionally grounded evaluation scenario (68 articles), which omits human involvement. Evaluation with users is relatively scarce, with 31 articles conducting human-grounded and nine papers performing an application-grounded evaluation. Compared to design-oriented research, behavioral science studies are rare (24%).

Analysis of XAI research areas in IS literature

To derive XAI research areas in IS literature, we identify patterns of homogenous groups of articles according to conceptual characteristics using cluster analysis. Cluster analysis is widely used in IS research as an analytical tool to classify and disentangle units in a specific context (Balijepally et al., 2011; Xiong et al., 2014) and to form homogenous groups of articles (Rissler et al., 2017; Xiong et al., 2014).

In our case, clustering is based on underlying XAI concepts and the methodological orientation of articles (cf. Fig. 4). To consider dimensions equally, we encoded articles

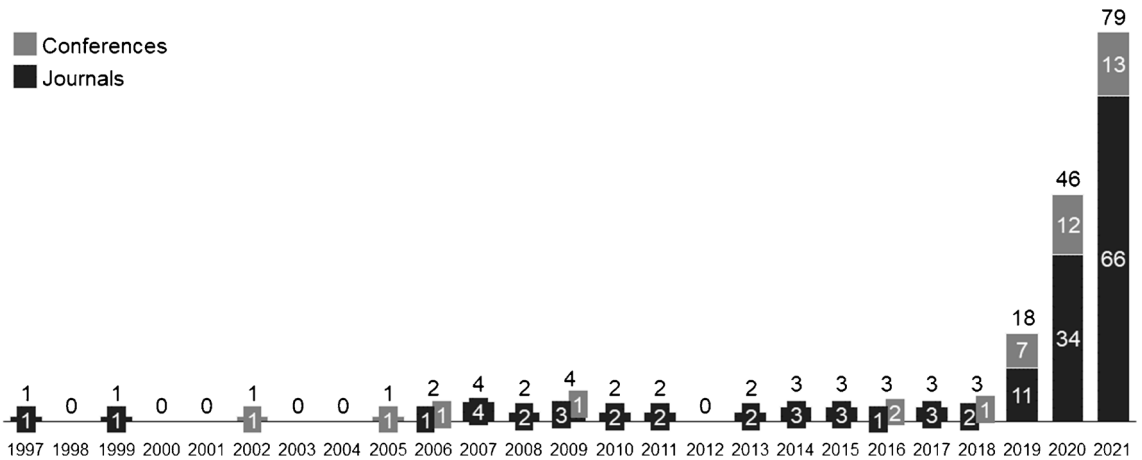
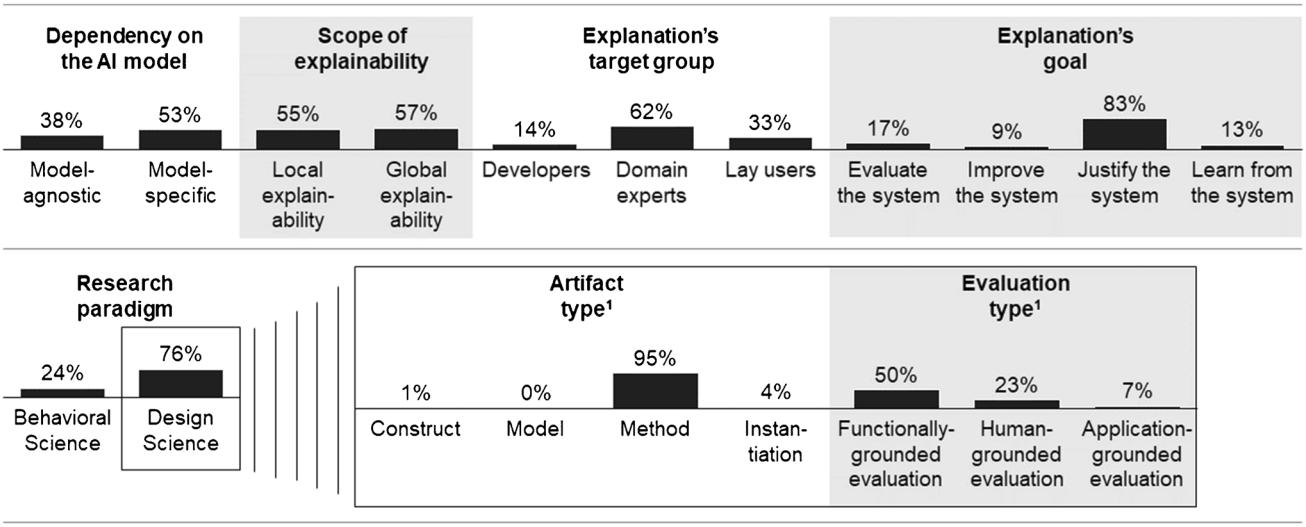


Fig. 3 Number of articles by year



1. Percentages refer to the 137 Design Science papers

Fig. 4 Characteristics of the academic discussion according to dimensions of the analysis scheme

as binary variables and normalized multiple answers per category. We applied the well-established agglomerative hierarchical clustering method using Euclidean distance measure as the similarity criterion and average linkage to group articles in clusters (Gronau & Moran, 2007). We chose this method as it does not form a predefined number of clusters but all possible clusters. To determine a reasonable number of clusters, we analyzed average silhouette scores (Shahapure & Nicholas, 2020). It resulted in eight clusters and two outliers with a positive average silhouette score (0.3), suggesting a solid clustering structure with an interpretable number of clusters.

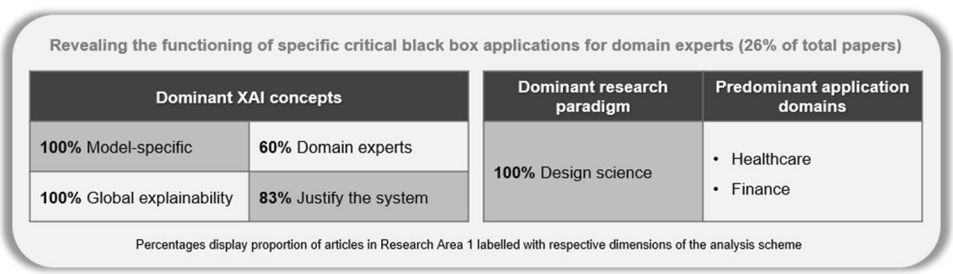
The clusters correspond to eight XAI research areas in IS literature, described in the following.

Research Area 1: Revealing the functioning of specific critical black box applications for domain experts

AI systems are increasingly applied in critical areas such as healthcare and finance, where there is a need for transparency in decision-making (He et al., 2006; Peñafiel et al., 2020; Pierrard et al., 2021). Transparency is meant to justify the usage of AI systems in such critical areas (Pessach et al., 2020). Research Area 1, which is among the largest with 47 papers (26%), aims at methods to reveal the functioning of specific critical black box applications to their users. For instance, XAI methods extract rules that reveal the functioning of an automatic diagnosis system to medical experts (Barakat et al., 2010; Seera & Lim, 2014) or, in the context of electronic markets, showcase central factors for loan approval on peer-to-peer lending platforms (Yang et al., 2021) (Fig. 5).

In critical application domains "where the cost of making a mistake is high" (Pierrard et al., 2021, p. 2), AI systems have the potential to serve as high-performant decision support systems—however, their lack of transparency constitutes a problem (e.g., Areosa & Torgo, 2019). To increase acceptance and adoption, researchers stress the need to justify their functioning to their users (Areosa & Torgo, 2019). For instance, medical practitioners not only need accurate predictions supporting their diagnosis but "would like to be convinced that the prediction is based

Fig. 5 Overview Research Area 1



on reasonable justifications" (Seera & Lim, 2014, p. 12). Thus, this research area aims at decision support systems that allow users to understand their functioning and predictive performance (Areosa & Torgo, 2019). To this end, explainable components are added to AI-based decision support systems for, e.g., diagnosis of diseases (Barakat et al., 2010; Singh et al., 2019; Stoean & Stoean, 2013), hiring decisions (Pessach et al., 2020), credit risk assessment (e.g., Florez-Lopez & Ramon-Jeronimo, 2015; Guo et al., 2021; Sachan et al., 2020), or fraud analysis in telecommunication networks (Irarrázaval et al., 2021). Studies in the healthcare domain identify that adding XAI methods for diagnosing diabetes increases medical accuracy and intelligibility by clinical practitioners (Barakat et al., 2010).

In Research Area 1, only very few articles develop XAI methods specifically for electronic markets or evaluate them in electronic markets. For instance, Nascita et al. (2021) develop a novel XAI approach for classifying traffic generated by mobile applications increasing the trustworthiness and interpretability of the AI system's outcomes. Grisci et al. (2021) evaluate their method for explaining neural networks on an online shopping dataset. They present a visual interpretation method that identifies which features are the most important for a neural network's prediction. While not explicitly designed for electronic markets, other methods might be transferable. Domain experts in electronic markets might benefit from global explanations, for instance, to improve supply chain management for B2B sales platforms or electronic purchasing systems.

Transparency of AI-based decision support systems is achieved by global explanations, which are supposed to reveal the functioning of the AI model as a whole rather than explain particular predictions (e.g., Areosa & Torgo, 2019; Pessach et al., 2020; Zeltner et al., 2021). Many approaches in Research Area 1 acquire a set of rules that approximate the functioning of an AI model (e.g., Aghaeipoor et al., 2021; Singh et al., 2019). For instance, researchers propose to produce explanatory rules in the form of decision trees from AI models to enable domain users such as medical practitioners to comprehend an AI system's prediction (Seera & Lim, 2014). More recently, approaches to approximate deep learning models with fuzzy rules have been pursued (e.g., Soares et al., 2021).

In an early paper, Taha and Ghosh (1999) emphasize the need to evaluate rule extraction approaches using fidelity, i.e., the capability to mimic the embedded knowledge in the underlying AI system. This is equivalent to functionally grounded evaluation, which is applied in many papers in Research Area 1 (62%). For instance, Soares et al. (2021) implement their rule extraction approach on several datasets and prove that it yields higher predictive accuracy than state-of-the-art approaches. Notably, only 6% of articles use users to evaluate explanations. For instance, Bresso et al. (2021) ask three pharmacology experts to evaluate whether extracted rules are explanatory for the AI system's outcomes, i.e., prognoses of adverse drug reactions. Irarrázaval et al. (2021) go further and perform an application-grounded evaluation. In a case study, they implement their explainable decision support system with a telecommunication provider and confirm that it helps reduce fraud losses. Thirty-four percent of papers demonstrate the technical feasibility of their methods and present how resulting explanations look like; however, they are not further evaluated.

Accordingly, a more robust evaluation, including users, may pave the way for future research in this research area, as suggested by Kim et al., (2020b). Other recurring themes of future research include the expansion of the developed ideas to other applications (Florez-Lopez & Ramon-Jeronimo, 2015; Sevastjanova et al., 2021). Finally, researchers often stress that explanations resulting from their approach are only one step toward a better understanding of the underlying AI system. Thus, it is essential to supplement and combine existing XAI approaches to help users gain a more comprehensive understanding (Murray et al., 2021).

Research Area 2: Revealing the functioning of specific black box applications for developers

The relatively small Research Area 2 consists of five papers (3%) and develops—similar to Research Area 1—methods to reveal the functioning of specific black box applications. Contrary to Research Area 1, which addresses domain experts, Research Area 2 focuses on explanations for developers. Explanations aim to provide insights into the functioning of opaque AI models to facilitate the development and implementation of AI systems (Martens et al., 2009) (Fig. 6).

Research Area 2 tackles the challenges of the growing complexity of AI models for developers: While predictions of more complex models often become more accurate, they also become less well understood by those implementing them (Eiras-Franco et al., 2019; Islam et al., 2020). Developers need information on how AI models process data and which patterns they discover to ensure that they are accurate and trustworthy (Eiras-Franco et al., 2019; Islam et al., 2020; Santana et al., 2007). Explanations can extract this information (Jakulin et al., 2005) and assist developers in validating a model before implementation, thereby improving its performance (Martens et al., 2009; Santana et al., 2007).

To this end, Research Area 2 develops model-specific XAI methods that generate global explanations and resemble those in Research Area 1. To illustrate, Martens et al. (2009) propose an approach to extract rules that represent the functioning of complex support vector machines (SVMs) and increase performance in predictive accuracy and comprehensibility. Eiras-Franco et al. (2019) propose an explainable

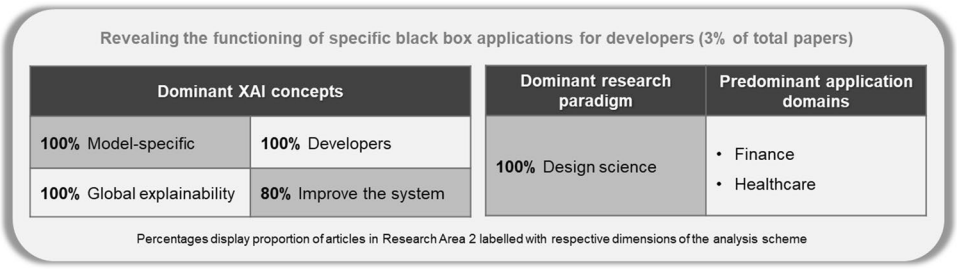


Fig. 6 Overview Research Area 2

method that improves both accuracy and explainability of predictions when describing interactions between two entities in a dyadic dataset. Due to the rather technical nature of the papers in Research Area 2, methods are not designed for or evaluated with electronic markets so far. However, XAI approaches in this research area might serve as a starting point to design novel XAI systems for digital platforms, for example, credit or sales platforms featuring AI systems.

Proof whether resulting explanations assist developers, as intended, is still pending. None of the papers in Research Area 2 includes an evaluation with humans. Sixty percent perform a functionally grounded evaluation. For instance, Martens et al. (2009) implement their rule extraction approach on several datasets and prove that it yields a performance increase in predictive accuracy compared to other rule extraction approaches.

The lack of evaluation with humans directly translates into a call for future research. In the next step, researchers should investigate the quality and efficacy of explanations from developers' perspectives. Moreover, in line with the rather technical focus of this research, improvements in the technical applicability of XAI methods, such as calculation speed, are suggested (Eiras-Franco et al., 2019).

Research Area 3: Explaining AI decisions of specific critical black box applications for domain experts

When utilizing complex AI systems as tools for decision-making, the reasons for particular AI outcomes often remain impenetrable to users. However, especially in critical application domains, AI decisions should not be acted upon blindly, as consequences can be severe (e.g., Gu et al., 2020; Su et al., 2021; Zhu et al., 2021). Thus, Research Area 3, encompassing 24 papers (13%), proposes XAI methods to generate explanations for particular outcomes of specific AI-based decision support systems. Decision support systems incorporating AI predictions and respective explanations serve to support domain experts in their daily work. Examples include anticipation of patient no-show behavior (Barrera Ferro et al., 2020), legal judgments (Zhong et al., 2019), and fault detection in industrial processes (Ragab et al., 2018). Some XAI methods are specifically designed for application in electronic markets, for example, mobile malware prediction (Iadarola et al., 2021), early risk detection in social media (Burdisso et al., 2019), and cost prediction for digital manufacturing platforms (Yoo & Kang, 2021) (Fig. 7).

Researchers commonly agree that AI-based decision support systems must be accompanied by explanations to effectively assist practitioners (e.g., Chatzimparmpas et al., 2020; Gu et al., 2020; Kwon et al., 2019). Thereby, explanations help practitioners better understand AI's reasoning, appropriately trust AI's recommendations, and take the best possible decisions (Hatwell et al., 2020; Hepenstal et al., 2021; Sun et al., 2021). Against this background, explanations are designed to be user-centric, i.e., to address the specific needs of certain (groups of) users. For instance, Barrera Ferro et al. (2020) propose a method to help healthcare professionals counteract low attendance behavior. Their XAI-based decision support system identifies variables explaining no-show probabilities. By adding explainability, the authors aim to prevent both practical and ethical issues when implementing the decision support system in a preventive medical care program for underserved communities in Columbia, identifying, e.g., income and local crime rates affect no-show probabilities.

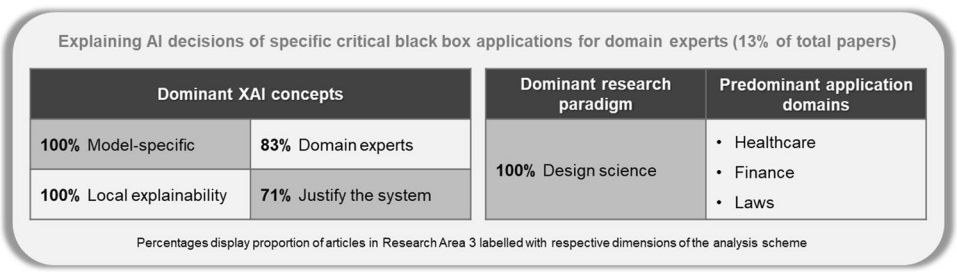


Fig. 7 Overview Research Area 3

To provide domain experts with explanations that meet their requirements, XAI methods to produce visual explanations along AI decisions are often employed: For instance, Gu et al. (2020) utilize an importance estimation network to produce visual interpretations for the diagnoses made by a classification network and demonstrate that the proposed method produces accurate diagnoses along fine-grained visual interpretations. Researchers argue that visualization allows users to easily and quickly observe patterns and test hypotheses (Kwon et al., 2019). Considering the drawbacks, visualizations of large and complex models such as random forests remain challenging (Neto & Paulovich, 2021).

Research Area 3 provides an above-average quota of evaluations with humans (33%). Majorly, researchers conduct user studies to assess the effectiveness of explanations (e.g., Chatzimparmpas et al., 2020; Neto & Paulovich, 2021; Zhao et al., 2019; Zhong et al., 2019). For example, Zhao et al. (2019) conduct a qualitative study with students and researchers to investigate the perceived effectiveness of an XAI-based decision support system in helping users understand random forest predictions in the context of financial scoring. Kumar et al. (2021) even go a step further and implement their XAI approaches in clinical practice to evaluate the trust level of oncologists working with a diagnosis system.

Existing research paves the way for three patterns with regard to future opportunities. First, researchers stress the need for other types of explainability to ensure a sufficient understanding of AI by users (Neto & Paulovich, 2021). Second, researchers propose to transfer XAI methods to different applications (Mensa et al., 2020). For instance, a novel XAI approach to design a conversational agent (Hepenstal et al., 2021) could also be applied in electronic markets. Third, whenever human evaluation is conducted in simulated scenarios with simplified tasks, there is a call to conduct application-grounded evaluation, such as field studies (Chatzimparmpas et al., 2020) and long-term studies (Kwon et al., 2019).

Research Area 4: Explaining AI decisions of specific black box applications for lay users

Similar to Research Area 3, Research Area 4, with seven papers (4%), focuses on model-specific XAI approaches to produce local explanations. While Research Area 3 targets AI users in a professional context, XAI approaches in Research Area 4 address lay people, such as users of a music platform seeking personalized recommendations (Kouki et al., 2020) or evaluating whether texts are similar in terms of meaning (Lopez-Gazpio et al., 2017). Thus, this research area is highly relevant for electronic markets (Fig. 8).

Given that AI finds its way to many areas of everyday life, the relevance of providing lay users with tailored support when faced with AI systems increases (Wang et al., 2019). The "target of XAI [in Research Area 4] is an end user who depends on decisions, recommendations, or actions produced by an AI and therefore needs to understand the rationale for the system's decisions" (Kim et al., 2021, p. 2). Often, lay users, such as people affected by automated AI decisions or users of AI in daily life, are assumed to provide a relatively low level of AI literacy (Wang et al., 2019). Explanations shall help them to easily scrutinize AI decisions and confidently employ AI systems (Kim et al., 2021; Kouki et al., 2020). Like in Research Area 3, researchers predominantly develop approaches to generate explanations for particular outcomes of specific AI models. Most resulting explanations are visual (Kim et al., 2021, 2020a; Kouki et al., 2020; Wang et al., 2019).

Research Area 4 provides an above-average percentage of evaluation with (potential) users (57%) (Kim et al., 2021; Kouki et al., 2020; Lopez-Gazpio et al., 2017). For instance, Kim et al. (2021) experimented with undergraduate students using an XAI system for video search to evaluate the quality of explanations and their effect on users' level of trust. They find that the XAI system yields a comparable level of efficiency and accuracy as its black box counterpart if the user exhibits a high level of trust in the AI explanations. Lopez-Gazpio et al. (2017) conduct two user studies to show that users perform AI-supported text processing tasks better with access to explanations. Only one paper follows functionally grounded evaluation, using a Netflix dataset (Zhdanov et al., 2021), showing that explainability does not need to impact predictive performance negatively.

One commonly mentioned avenue for future research is to transfer XAI approaches—which are often developed for specific applications—to other contexts. For instance, an XAI approach designed for a medical diagnosis tool for lay users might also be beneficial when integrated into a fitness

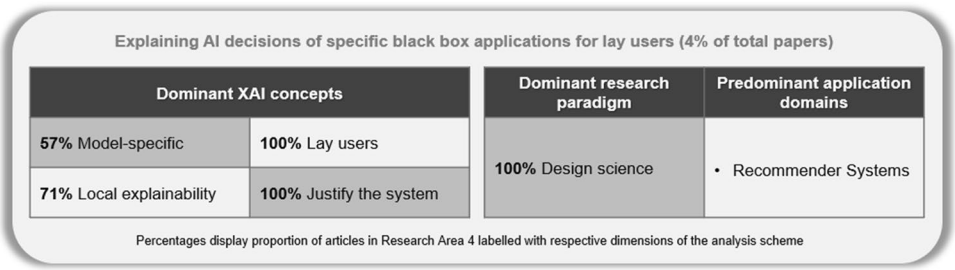


Fig. 8 Overview Research Area 4

app (Wang et al., 2019). While the authors formulate the need to investigate the effectiveness of explanations for lay users (Kouki et al., 2020), the lack of functionally grounded evaluation also translates into a call for a technical assessment and improvement of XAI approaches, such as computation time (Kim et al., 2020a).

Research Area 5: Explaining decisions and functioning of arbitrary black boxes

The ubiquitous nature of AI and its deployment in an increasing variety of applications is accompanied by a rising number of AI models. Consequently, the need for XAI approaches that can work independently from the underlying AI model arises (e.g., Ming et al., 2019). Research Area 5, among the most prominent research areas with 52 papers (29%), addresses this call and develops model-agnostic XAI approaches (Moreira et al., 2021). Many methods have already been applied for electronic markets, for example, for B2B sales forecasting (Bohanec et al., 2017) or prediction of Bitcoin prices (Giudici & Raffinetti, 2021) (Fig. 9).

Papers in Research Area 5 are also driven by the desire to make the outcomes and functioning of AI systems more understandable to users (Fernandez et al., 2019; Li et al., 2021; Ribeiro et al., 2016). First and foremost, explanations intend to assist users in appropriately trusting AI, i.e., critically reflecting on an AI system's decision instead of refusing to use it or blindly following it (Förster et al., 2020b). However, aiming to contribute to the explainability of arbitrary AI models, methods differ from Research Areas 1 to 4 in two ways.

First, methods are not designed to address specific needs in certain applications but aim to explain how and why models make their decisions in general (e.g., Blanco-Justicia et al., 2020). The target group are users of all "domains where ethical treatment of data is required" (Ming et al., 2019, p. 1), including domain experts (79%), such as managers or decision-makers (Bohanec et al., 2017) as well as lay users (38%), such as social media users supported by AI to detect hate speech (Bunde, 2021). In the latter example, researchers show that a dashboard showing and explaining whether a text contains hate is perceived as valuable by users, and that the XAI feature increased the perception of usefulness, ease of use, trustworthiness, and use intention of the artifact. Explanations are constructed to address the standard requirements of various AI users of different application domains. As a result, explanations are often accessible to a wider audience and help users with little AI experience understand, explore, and validate opaque systems (Ming et al., 2019). For example, for the identification of diseases on an automatic diagnosis platform for doctors and patients, building an understandable diagnostics flow for doctors and patients (Zhang et al., 2018). Second, the XAI methods are not designed to be technically tied to specific AI models, but to be applied to various AI models (Mehdiyev & Fettke, 2020, p. 4). Thus, XAI approaches in this area only access the inputs and outcomes without making architectural assumptions regarding the AI model (Ming et al., 2019).

Most papers in Research Area 5 focus on local explanations (73%). A well-known local method is LIME which identifies important features for particular AI predictions by learning easy-to-interpret models locally around the inputs (Ribeiro et al., 2016). Researchers stress that explanations should be human-friendly to facilitate human understanding of the reasons for AI decisions (e.g., Cheng et al., 2021). For instance, Blanco-Justicia et al. (2020) aim at human-comprehensible explanations by limiting the depth of decision trees that approximate the AI model's functioning. Many researchers focus on methods to generate counterfactual explanations, which align with how humans construct explanations themselves (Cheng et al., 2021; Fernandez et al., 2019; Förster et al., 2021). Counterfactual explanations point out why the AI system yields a particular outcome instead of another similarly perceivable one.

The focus of Research Area 5 lies on the XAI methods themselves rather than specific applications. Accordingly, researchers choose relevant but exemplary use cases to evaluate their proposed XAI methods, such as the prediction of credit risk (Bastos & Matos, 2021), churn prediction (Lukyanenko et al., 2020), or mortality in intensive care units (Kline et al., 2020). To demonstrate versatile applicability, researchers often implement their approaches on a range of datasets from different domains including applications in electronic markets such as fraud detection (Hardt et al., 2021) or news-story classification for online advertisements, which helps improve data quality and model performance

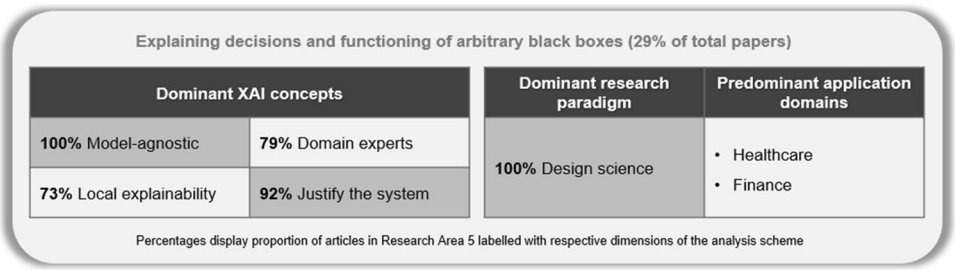


Fig. 9 Overview Research Area 5

(Martens & Provost, 2014). XAI approaches in Research Area 5 could beyond be applied to electronic markets—for example, an XAI dashboard consolidating a large amount of data necessary for child welfare screening is also considered helpful for different data-intensive online platforms (Zytek et al., 2021).

Like in Research Areas 1 and 2, most papers conduct functionally grounded evaluation (52%). However, as repeatedly stated by the authors in this research area, XAI methods are designed to assist humans in building appropriate trust (e.g., Bunde, 2021; van der Waa et al., 2020). Accordingly, in recent years, papers include evaluations with users (46%) (Abdul et al., 2020; Hardt et al., 2021; Ming et al., 2019). User studies serve, for instance, to assess perceived characteristics of explanations (Förster et al., 2020b, 2021) or to compare the utility of different explanations for decision-making (van der Waa et al., 2020). Researchers often resort to simplified tasks with subjects being students (Štrumbelj & Kononenko, 2014) or recruited via platforms like Amazon Mechanical Turk (van der Waa et al., 2020).

As evaluation is often conducted in somewhat artificial settings, researchers propose to evaluate model-agnostic XAI methods in realistic or real settings, for instance, through field experiments (Bohanec et al., 2017; Förster et al., 2020b, 2021; Giudici & Raffinetti, 2021). Other recurring themes for future research include the expansion of the ideas to other application domains (e.g., Spinner et al., 2020; Zytek et al., 2021). Finally, further empirical research is requested to identify required modifications of existing XAI approaches and specific requirements that can serve as a starting point for the design of novel XAI methods (Moradi & Samwald, 2021).

Research Area 6: Investigating the impact of explanations on lay users

There is a substantial body of literature developing XAI methods to automatically generate explanations (cf. Research Areas 1 to 5); however, insights on the role of explainability in human-AI interaction are somewhat rare (Ha et al., 2022; Narayanan et al., 2018; Schmidt et al., 2020). Against this background, this research area with 22 articles (12%) empirically investigates user experience and user behavior in response to explanations, such as understanding of and trust in the underlying AI system (Dodge et al., 2018; Shin, 2021a; van der Waa et al., 2021). The focus lies on lay users as an explanation's target group of (100%). Many papers investigate XAI for electronic market applications—for example, recommendation of online news articles (Shin, 2021a), intelligent tutoring (Conati et al., 2021), or credit risk assessment (Moscato et al., 2021) (Fig. 10).

Researchers stress the importance of involving users to derive how explanations should be designed (Wanner et al., 2020b). Articles in this research area pursue two goals: (i) generating insights on how explanations affect the interaction between users and AI and (ii) deriving requirements for adequate explanations. More concretely, researchers investigate lay user experience and lay user behavior, such as trust (Alam & Mueller, 2021; Burkart et al., 2021; Conati et al., 2021; Hamm et al., 2021; Jussupow et al., 2021; Schmidt et al., 2020; Shin, 2021a, 2021b), understanding (Lim et al., 2009; Shen et al., 2020; Shin, 2021a, 2021b; van der Waa et al., 2021), perception (Fleiß et al., 2020; Ha et al., 2022; Jussupow et al., 2021; Shin, 2021a), and task performance (van der Waa et al., 2021). Lay users considered are, for instance, potential job candidates interacting with conversational agents in recruiting processes (Fleiß et al., 2020) or diabetes patients interacting with a decision support system to determine the correct dosage of insulin (van der Waa et al., 2021). Based on their findings, researchers contribute knowledge on how practical explanations can be designed (Dodge et al., 2018; Förster et al., 2020a; Wanner et al. 2020b). Most of these findings are valid for electronic markets, such as AI-led moderation for eSports communities (Kou & Gui, 2020) or patient platforms with AI as the first point of contact (Alam & Mueller, 2021). The authors of the latter study find that visual and example-based explanations had a significantly better impact on patient satisfaction and trust than text-based explanations or no explanations at all.

A reoccurring study design to investigate user experience and behavior is a controlled experiment with human subjects performing simplified tasks (Lim et al., 2009). For example, Burkart et al. (2021) investigate users' willingness to adapt their initial prediction in response to four treatments with different degrees of explainability. Surprisingly, in their

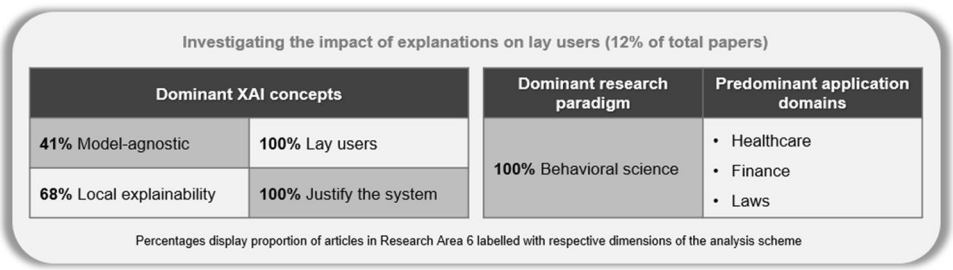


Fig. 10 Overview Research Area 6

specific study, all participants improved their predictions after receiving advice, regardless of whether it featured an explanation. Likewise, Jussupow et al. (2021) investigate users' trust in a biased AI system depending on whether explanations are provided or not. They find that users with low awareness of gender biases perceive a gender-biased AI system that features explanations as trustworthy, as it is more transparent than a system without explanations. Focusing on user experience, Shen et al. (2020) examine users' subjective preferences for different degrees of explainability. Only a few papers build their work on existing theories. For instance, Hamm et al. (2021) adapt the technology acceptance model to examine the role of explainability on user behavior.

The results in Research Area 6 reveal that explanations indeed affect user experience and user behavior. Most papers propose a positive effect on human-AI interaction, such as an increase of users' trust in the AI system (Lim et al., 2009) or intention to reuse the system (Conati et al., 2021). However, some studies indicate a contrary effect, i.e., participants supported by an AI-based decision support tool for text classification reported reduced trust in response to increased transparency (Schmidt et al., 2020). Beyond, the findings of this research area inform how explanations should be built to be effective. For instance, Burkart et al. (2021) found that while local and global explanations help improve participants' decisions, local explanations are used more often. The findings by Förster et al. (2020a) indicate that concreteness, coherence, and relevance are decisive characteristics of local explanations and should guide the development of novel XAI methods. Overall, researchers conclude that user involvement is indispensable to assess if researchers' assumptions on explanations hold (Shin, 2021a; van der Waa et al., 2021).

Results from this research area mainly stem from experiments with recruited participants for simplified tasks, such as students (Alam & Mueller, 2021). Paving the way for future research, researchers stress the importance of verifying findings with real users performing actual tasks (Shen et al., 2020). Furthermore, there is a call for longitudinal studies considering that users' characteristics and attitudes might change over time (Shin, 2021a). Finally, while first progress is made to consider mediating factors predicting the influence of explainability (e.g., Shin, 2021a), most works do not tie their studies to theories; thus, there is a call for developing and testing theories (Hamm et al., 2021).

Research Area 7: Investigating the impact of explanations on domain experts

Most XAI methods are designed to assist domain experts in interacting with AI-based decision support systems. To better understand how explainability influences user experience and user behavior in this regard, Research Area 7 includes 17 empirical papers (9%) with a focus on domain experts, such as doctors (Ganeshkumar et al., 2021; Kim et al., 2020b) or decision-makers in credit scoring (Huysmans et al., 2011). Compared to Research Area 6, fewer papers investigate the impact of explanations in electronic market applications. Examples include an AI-based scheduling platform for healthcare professionals (Schlicker et al., 2021) and an AI web application for patient analysis and risk prediction (Fang et al., 2021) (Fig. 11).

Researchers argue that while there is agreement on the need to increase the explainability of critical AI applications, insights on how different explanation types affect the interaction of domain experts with AI is rare (Liao et al., 2020). This research area aims to understand the impact of explainability concerning user experience and user behavior in the context of AI-based decision support systems (Chakraborty et al., 2021; Elshawi et al., 2019; Liao et al., 2020; Martens et al., 2007). Similar to Research Area 6, findings aim to provide knowledge on how to design adequate explanations, however, with a focus on domain experts (Liao et al., 2020; Wanner et al., 2020a).

A reccurring research approach is to conduct experiments investigating the impact of explainability on users' decision-making with AI. In a pioneering paper, Huysmans et al. (2011) examine how different degrees of explainability affect AI system comprehensibility in a laboratory experiment. They find that decision tables perform significantly better than decision trees, propositional rules, and oblique rules with regard to accuracy, response time, answer confidence, and ease of use. Moreover, researchers conduct interviews to assess user needs for explainability in critical AI applications (Liao et al., 2020).

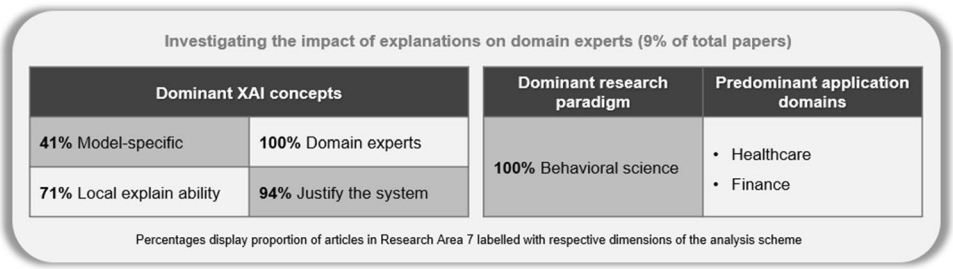


Fig. 11 Overview Research Area 7

Overall, findings from these studies indicate that explainability can positively influence user experience and user behavior of domain experts. The findings by Huysmans et al. (2011) outlined above suggest that explainability in the form of decision tables can lead to faster decisions while increasing answer confidence. Additionally, findings inform how explanations should be designed and applied to yield specific effects. For example, Elshawi et al. (2019) reveal that local explanations are suitable for medical diagnoses to foster users' understanding while global explanations increase users' understanding of the entire AI model. Although this research area proves the benefit of XAI for domain experts, practitioners still struggle with the gaps between existing XAI algorithmic work and the aspiration to create human-consumable explanations (Liao et al., 2020).

While existing studies show that types of explanations, such as local and global explanations, vary in effectiveness on users' system understanding, future research may deepen these insights and investigate other concepts, such as concreteness and coherence. Furthermore, researchers stress the importance of further investigating how users' characteristics moderate explanations' influence on user experience and user behavior (Bruijn et al., 2021). Expert users of electronic markets are not the focus of research attention yet. Finally, while most researchers focus on the impact of explanations on users' perceptions and intentions, there is a call for research on actual behavior (Bayer et al., 2021).

Research Area 8: Investigating employment of XAI in practice

In contrast to Research Areas 6 and 7, which comprise empirical studies to investigate user experience and user behavior, Research Area 8 focuses on technical and managerial aspects of XAI in practice. For instance, researchers conduct case studies to examine scalability (Sharma et al., 2020) and trade-offs of XAI in practice (Tabankov & Möhlmann, 2021). The four papers (2%), which all were published between 2019 and 2021, represent the smallest research area. Findings predominantly address developers (100%) and managers who want to implement XAI in organizations (Sharma et al., 2020) (Fig. 12).

The motivation for this research area is a scarce understanding of organizational and technical challenges practitioners face when implementing explanations for AI (Hong et al., 2020). Researchers agree that this might hinder XAI from addressing critical real-world needs. Against this background, empirical studies aim to generate insights into how XAI can be successfully employed in organizations (Hong et al., 2020; Tabankov & Möhlmann, 2021).

To this end, Hong et al. (2020) conduct semi-structured interviews with industry practitioners to examine the role of explainability when developers plan, build, and use AI models. One important finding is the high practical relevance of scalability and integrability of XAI methods—which has not yet been the focus of existing research. Building on these insights, Sharma et al. (2020) evaluate the performance of XAI methods with respect to technical aspects in an electronic market–related case study, i.e., anomaly detection for cloud-computing platforms. Findings reveal that the computation time of tree-based XAI methods should be improved to enable the large-scale application. Tabankov and Möhlmann (2021), with their case study, take a managerial perspective and investigate trade-offs between explainability and accuracy of XAI for in-flight services. Findings suggest that compromises and limitations for both sides have to be weighed during the implementation process.

Insights from this research area pave the way for future research: First, when developing novel XAI methods, researchers should consider technical aspects, first and foremost, scalability (Hong et al., 2020). This is especially relevant for electronic market applications, which often need to adapt to sudden user growth. Second, more empirical research on XAI from an organizational and managerial perspective is needed. In particular, further research might provide deeper insights into whether and to what extent explainability is needed to achieve organizational goals (Tabankov & Möhlmann, 2021). Third, there is a call for insights into the demands of XAI developers (Hong et al., 2020).

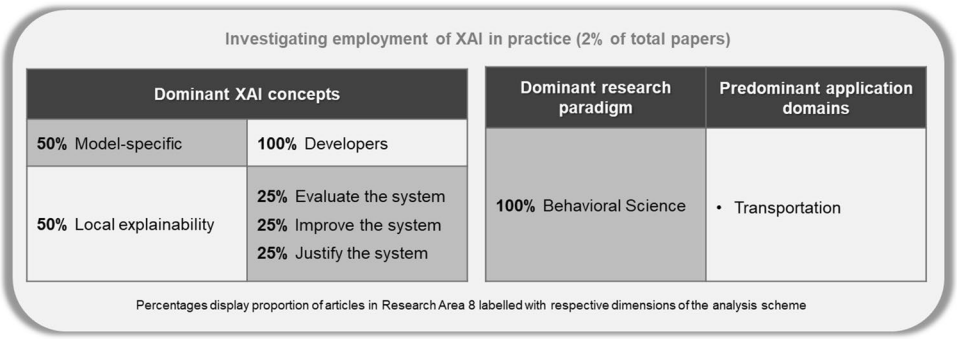


Fig. 12 Overview Research Area 8

Synthesis of XAI research areas in IS literature

In sum, based on theoretical concepts of XAI research and methodological concepts of IS research, a cluster analysis reveals eight major XAI research areas in IS literature (cf. Fig. 13, Appendix).

Five research areas (76% of all papers in our corpus) deal with developing novel XAI approaches. This body of literature can be further differentiated depending on the underlying XAI concepts, first and foremost dependency on the AI model and scope of explainability, as well as whom explanations address. Research Area 1 and Research Area 2 both focus on model-specific XAI approaches to generate global explanations for expert audiences—domain experts in Research Area 1, and developers in Research Area 2. Research Area 3 and Research Area 4 entail largely local explanations for specific AI models that address domain experts and lay users, respectively. Research Area 5 features model-agnostic approaches. Overall, the primary purpose of explanations is to justify the (decisions of) AI systems (Research Areas 1, 3, 4, and 5).

The remaining three research areas comprise fewer articles (24%) focusing on behavioral science research. Note that in our case, the term "behavioral science" not only refers to studies that build and justify theory, for instance, in deriving and testing hypotheses but, more generally, includes research that aims at generating empirical insights. Indeed, only a few XAI papers in IS derive and test hypotheses. Empirical research in our corpus can be distinguished by its focus on specific target groups. While Research Area 6 focuses on lay users, Research Area 7 deals with users with domain knowledge. Research Area 8 focuses on developers.

Discussion and conclusion

We conducted a systematic and structured review of research on XAI in IS literature. This section outlines opportunities for future research that may yield interesting insights into the field but have not been covered so far. Subsequently, we describe our work's contribution, implications, and limitations.

Future research agenda

Our synthesis reveals five overarching future research directions related to XAI research in IS, which, along with a related future research agenda, are outlined below: (1) refine the understanding of XAI user needs, (2) reach a more comprehensive understanding of AI, (3) perform a more diverse mix of XAI evaluation, (4) solidify theoretical foundations on the role of XAI for human-AI interaction, and (5) increase and improve the application to electronic market needs. Note that the future research directions and future research agenda are by no means exhaustive but intend to highlight and illustrate potential avenues that seem particularly promising.

Future Research Direction 1: Refine the understanding of XAI user needs

XAI research is criticized for not focusing on user needs, which is a prerequisite for the effectiveness of explanations (cf. Herse et al., 2018; Meske et al., 2020). Indeed, as argued in many papers in the different research areas identified, there is still a gap between the research's focus on novel algorithms and the aspiration to create human-consumable explanations (e.g., Liao et al., 2020; Seera & Lim, 2014). Areosa and Torgo

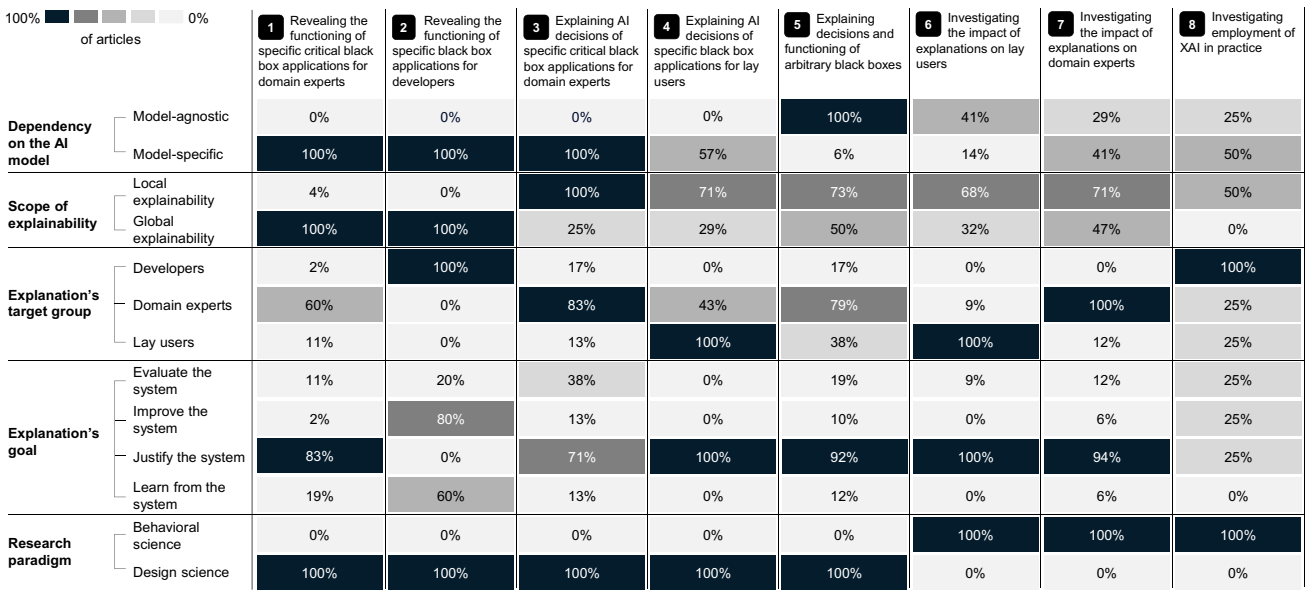


Fig. 13 Synthesis of XAI research areas in IS literature

(2019) stress the necessity to provide insights into the type of usage and information XAI tools bring to end users. As one of the foci in IS research is the design of user-centric and interactive technologies, IS research is predestined to put the user at the center of attention and make explanations understandable (Bauer et al., 2021). While six of the eight research areas focus on broader user groups, i.e., lay users, domain experts, or developers, only a few studies base the design of XAI approaches on specified target users and their needs (e.g., medical experts with different level of domain knowledge). This shortcoming has already been raised in studies that call for a more user-specific design of XAI solutions (cf. Abdul et al., 2018; Miller, 2019). However, only a few studies have implemented user-specific designs so far. For instance, Barda et al. (2020) propose an XAI approach that produces explanations for predictions based on a pediatric intensive care unit's mortality risk model. It considers user-specific explanation and information goals, which vary according to the clinical role (e.g., nurses and physicians). Further empirical insights highlight the necessity for the user-specific design of explanations, as XAI can only create human agency and appropriate trust if it considers the specific user needs (Dodge et al., 2018; Elshawi et al., 2019).

We identify several research opportunities to pave the way for a refined understanding of XAI user needs: First, more empirical research might sharpen insights into how different types of explanations affect the behavior and experience of various user groups and which effects different explanation types might have on these groups—for example, medical practitioners (e.g., Seera & Lim, 2014). Second, future research could refine the differentiation between developers, domain experts, and lay users, as other user characteristics besides expertise might play a central role (e.g., Cui et al., 2019). For instance, the user's knowledge structure, beliefs, interests, expectations, preferences, and personality could be considered (Miller et al., 2017). Third, the conjunction of user characteristics and the purpose of explanations could be analyzed, especially given that the purpose of explanations depends on the context and user type (Liao et al., 2020). Fourth, future research could put more emphasis on investigating the concrete XAI needs of developers, which would benefit from explainability (cf. Kim et al., 2021) but are so far seldomly addressed. This is underlined by the fact that in Research Area 2 ("Revealing the functioning of specific black box applications for developers"), the only research area focusing on developers, none of the papers evaluates its concepts with actual developers.

Future Research Direction 2: Reach a more comprehensive understanding of AI

While a plethora of techniques produce various types of explanations, only a few researchers combine different XAI approaches with the aim of a comprehensive understanding of AI. The overarching goal of XAI is to make AI systems and their outcomes understandable to humans, especially important when AI supports decision-making in critical areas such as healthcare and finance (Pessach et al., 2020). Single (types of) explanations are often insufficient to reach the ambitious goal of comprehensive user understanding. Many researchers underpin that their approaches are only one step toward a better understanding of the underlying AI systems (e.g., Moradi & Samwald, 2021; Neto & Paulovich, 2021). However, the question of how to synthesize different research efforts to get closer to a comprehensive understanding of AI systems has received little research attention. Especially in Research Area 1 ("Revealing the functioning of specific critical black box applications for domain experts") and Research Area 3 ("Explaining AI decisions of specific critical black box applications for domain experts"), both of which focus on domain experts, researchers identify the need for further explanation types to ensure that users can reach a more comprehensive understanding of AI (e.g., Murray et al., 2021; Neto & Paulovich, 2021).

Against this backdrop, promising future research opportunities arise: First, it could be beneficial to investigate the combination of different types of explanations which might complement each other for user understanding, e.g., local and global explanations, a call made in many of the analyzed papers (cf. Burkart et al., 2021; Elshawi et al., 2019; Mombini et al., 2021). So far, efforts on developing novel approaches mainly concentrate on either type, with only 18% of the papers combining local and global interpretability (e.g., Burkart et al., 2021; Elshawi et al., 2019). Second, a stronger focus on user interfaces might serve as an auspicious starting point for a more complete understanding of AI. For example, interactivity would allow users to explore an algorithm's behavior, and XAI approaches to adapt explanations to users' needs (Cheng et al., 2019). Ming et al. (2019) provide the first promising attempts in this direction, developing an interactive visualization technique to help users with little AI expertise understand, explore, and validate predictive models. Third, personalized explanations taking into account users' mental models and the application domain can foster understanding (Schneider & Handali, 2019). Kouki et al. (2020) are among the first to study the problem of generating and visualizing personalized explanations for recommender systems.

Future Research Direction 3: Perform a more diverse mix of XAI evaluation

Our analysis shows that existing IS literature on XAI exposes a one-sided tendency toward the functional evaluation of XAI approaches. Seminal design science contributions emphasize the need for rigor in evaluating IT artifacts, including

functional evaluations but also "the complications of human and social difficulties of adoption and use" (Venable et al., 2016, p. 82). While the latter plays a significant role in the context of XAI, 71% of the articles that develop XAI approaches in our corpus neglect evaluation with (potential) users. Only 6% combine functional evaluation with user evaluation. Thus, existing research runs the risk of inaccurate insights derived from unduly simplified evaluation scenarios (Wang et al., 2019). In almost all research areas, papers identify a better mix of evaluation methods as one of the most important directions for future research (e.g., Chatzimparmpas et al., 2020; Kim et al., 2020b).

Proposed avenues for further research are closely linked to a call for a more diverse mix of different kinds of evaluations (cf. Venable et al., 2016). First, XAI approaches should be more frequently evaluated with humans (cf. human-grounded evaluation) to take into account human risks associated with novel XAI approaches. For example, many papers in Research Area 1 ("Revealing the functioning of specific critical black box applications for domain experts") call for a more robust evaluation, including human users (e.g., Areosa & Torgo, 2019; Kim et al., 2020b). Second, there should be a stronger focus on evaluation with real users in real settings (cf. application-grounded evaluation) to assess the utility, quality, and efficacy of novel approaches in real-life scenarios. This point is stressed by several papers in Research Area 3 ("Explaining AI decisions of specific critical black box applications for domain experts") (e.g., Chatzimparmpas et al., 2020; Kwon et al., 2019) and Research Area 6 ("Investigating the impact of explanations on lay users") (e.g., Shen et al., 2020; Shin, 2021a). Third, novel evaluation strategies might be investigated that combine functionally and human-grounded evaluation to consolidate the benefits of both, i.e., the possibility of a robust comparison of competing XAI approaches at relatively low cost and the consideration of social intricacies.

Future Research Direction 4: Solidify theoretical foundations on the role of XAI for human-AI interaction

Our examination shows that XAI in IS research is predominantly not very theory-rich. While broad efforts to develop novel artifacts exist, only few papers (24%) explicitly focus on contributions to theory by conducting empirical research. These studies generate first exciting insights into how explainability may affect the experience and behavior of AI users (cf. Research Areas 6 and 7); however, only 13 papers explicitly tie their research to theory. The following IS theories have been used to investigate XAI in our literature corpus: Activity Theory (Kou & Gui, 2020), Agency Theory (Wanner et al., 2020a), Attribution Theory (Ha et al., 2022; Schlicker et al., 2021), Cognitive Fit Theory (Huysmans et al., 2011), Elaboration Likelihood Model/Heuristic Systematic Model (Shin, 2021a, 2021b; Springer & Whittaker, 2020), Information Boundary Theory (Yan & Xu, 2021), Information Foraging Theory (Dodge et al., 2018), Information Processing Theory (Sultana & Nemati, 2021), Psychological Contract Violation (Jussupow et al., 2021), Technology Acceptance Model/Theory of Planned Behavior/Theory of Reasoned Action (Bayer et al., 2021; Wanner et al., 2020a), Theory of Swift Trust (Yan & Xu, 2021), and Transaction Cost Theory (Wanner et al., 2020a). Mainly, cognitive theories are employed. As the human side of explanations is both social and cognitive, literature points out that explainability in the context of human-AI interaction should be viewed through a cognitive and a social lens (Berente et al., 2021; Malle, 2006). The extant studies pave the way for a diverse and meaningful XAI research agenda. It is crucial to add theoretical lenses (Wang et al., 2019), to deepen the understanding of the role of XAI for human-AI interaction. Extant literature stresses the need to further develop and test theories, for example, concerning the relationship between XAI and use behavior (Hamm et al., 2021).

Pursuing this avenue, first, we call to supplement insights based on cognitive theories by investigating XAI through a social lens. Second, it might be helpful not only to include and test IS theories but also theories from disciplines such as social sciences, management, and computer science. XAI is multidisciplinary by nature with people, information technology, and organizational contexts being intertwined. For instance, the social sciences might be promising to model user experience and behavior as they aim to understand how humans behave when explaining to each other (Miller, 2019). Third, as extant empirical studies are mostly limited to one-time interactions between humans and XAI, more research on the long-term influence of explanations is needed. For instance, the question of how explanations may sustainably change users' mental models and behavior should gain more attention. Papers in our body of literature also call for longitudinal studies considering that users' characteristics and attitudes might change over time (Shin, 2021a). Fourth, the organizational perspective on XAI is mainly neglected. Existing literature examines AI's influence on the competitiveness of companies (e.g., Rana et al., 2022). For different organizations, AI has become an essential source of decision support (Arrieta et al., 2020); thus, XAI is of utmost importance for bias mitigation (Akter et al., 2021a; Zhang et al., 2022). Therefore, it would be beneficial to examine the role of XAI from an organizational perspective as well.

Future Research Direction 5: Increase and improve the application to electronic market needs

The literature review shows that only a minority of extant studies aim at solving electronic market-related challenges (e.g., Burdisso et al., 2019; Irarrázaval et al.,

2021). Among business applications, XAI is especially relevant for electronic markets, as trust is paramount in all buyer-seller relationships (Bauer et al., 2020; Marella et al., 2020). Promising first studies on XAI in electronic markets focus on recurring use cases, for example, recommender systems in entertainment (e.g., Zhdanov et al., 2021), patient platforms in healthcare (e.g., van der Waa et al., 2021), and credit platforms in finance (e.g., Moscato et al., 2021). Given that electronic markets are increasingly augmented with AI-based systems and their complex nature is often an obstacle (Adam et al., 2021; Thiebes et al., 2021), electronic markets provide large potential for XAI research. To illustrate, the benefit of XAI could be explored for AI-based communication with customers on company platforms or AI-augmented enterprise IS for domain experts in supply chain or customer relationship management. While the benefits of XAI in electronic markets become obvious, an XAI research agenda with a focus on the needs of electronic markets might, in turn, benefit from diverse cases, including a variety of users.

There are three possible pathways in which researchers could address this issue and improve the application to electronic markets: First, existing XAI approaches could be transferred to and investigated in electronic markets. For instance, an XAI approach for conversational agents (Hepenstal et al., 2021) could be applied in electronic markets, for example, in the context of B2C sales platforms or for customer support. Second, given the strong interaction of people and technology in electronic markets (cf. Thiebes et al., 2021), it is pivotal to gain a better understanding of users' needs regarding the explainability of AI in electronic markets, for example, users of music platforms (Kouki et al., 2020), news websites (Shin, 2021a), or streaming platforms (Zhdanov et al., 2021) seeking personalized recommendations. Third, researchers could develop novel XAI methods and user interfaces that specifically meet electronic market needs, for instance, the ability to work with large amounts of data and provide interactive interfaces for business and private users. Table 2 summarizes the future research directions and opportunities outlined above.

Contribution

The contribution of our study is twofold. First, we provide a structured and comprehensive literature review of XAI research in IS. A literature review is especially important for a young and emerging research field like XAI, as it "uncover[s] the sources relevant to a topic under study" (vom Brocke et al., 2009, p. 13) and "creates a firm foundation for advancing knowledge" (Webster & Watson, 2002, p. 13). XAI draws from various scientific disciplines such as computer science, social sciences, and IS. While existing research already views XAI through the lenses of adjacent disciplines like social sciences (e.g., Miller, 2019), we accumulate the state of knowledge on XAI from the IS perspective. With its multiperspective view, IS research is predestined to investigate and design the explainability of AI. In turn, XAI can significantly contribute to the ongoing discussion of human-AI interaction in the IS

Table 2 Future research agenda

Future research directions	Future research opportunities
1: Refine the understanding of XAI user needs	• Pursue empirical research to sharpen understanding of how explanations affect behavior and experience of user groups • Refine differentiation between user groups for a more complete understanding of XAI end-user characteristics • Analyze the conjunction of XAI user characteristics and the purpose of explanations • Investigate the needs of developers in the context of XAI
2: Reach a more comprehensive understanding of AI	• Investigate the combination of different types of explanations • Investigate user interfaces with a focus on interactivity • Pursue personalized explanations taking users' mental models into account
3: Perform a more diverse mix of XAI evaluation	• Pursue evaluations with human users • Pursue evaluations with real users in real-life scenarios • Combine functionally and human-grounded evaluation
4: Solidify theoretical foundations on the role of XAI for human-AI interaction	• Investigate XAI through a social lens • Pursue interdisciplinary approaches, e.g., employ theories from the social sciences • Research the long-term influence of explanations, e.g., on users' mental models • Examine the role of XAI from an organizational perspective
5: Increase and improve the application to electronic market needs	• Transfer existing XAI approaches to electronic markets • Investigate user needs regarding the explainability of AI in electronic markets • Design XAI approaches that meet specific electronic market requirements

discipline. Compared to existing works on XAI in IS (e.g., Meske et al., 2020), our study is the first to synthesize XAI research in IS based on a structured and comprehensive literature search. The structured and comprehensive literature search reveals 180 research articles published in IS journals and conference proceedings. From 2019 onward, the number of published articles increased rapidly, resulting in 79% of the articles published between 2019 and 2021. Model-specific XAI methods (53%) are more often in focus than model-agnostic XAI methods (38%). Most articles address domain experts as the target group (62%) and focus on the justification of AI systems' decisions as XAI goal (83%). Extant IS research efforts concentrate on developing novel XAI artifacts (76%); however, only 23% of the proposed artifacts are evaluated with humans. A minority of studies aim at building and justifying theories or generating empirical insights (24%). Building on established XAI concepts and methodological orientation in IS, we are the first to derive XAI research areas in IS. Extant XAI research in IS can be synthesized in eight research areas: (1) Revealing the functioning of specific critical black box applications for domain experts (26% of papers), (2) Revealing the functioning of specific black box applications for developers (3% of papers), (3) Explaining AI decisions of specific critical black box applications for domain experts (13% of papers), (4) Explaining AI decisions of specific black box applications for lay users (4% of papers), (5) Explaining decisions and functioning of arbitrary black boxes (29% of papers), (6) Investigating the impact of explanations on lay users (12% of papers), (7) Investigating the impact of explanations on domain experts (9% of papers), (8) Investigating employment of XAI in practice (2% of papers).

Second, we provide a future research agenda for XAI research in IS. The research agenda comprises promising avenues for future research raised in existing contributions or derived from our synthesis. From an IS perspective, the following directions for future research might provide exciting insights into the field of XAI but have not yet been covered sufficiently: (1) Refine the understanding of XAI user needs, (2) Reach a more comprehensive understanding of AI, (3) Perform a more diverse mix of XAI evaluation, (4) Solidify theoretical foundations on the role of XAI for human-AI interaction, (5) Increase and improve the application to electronic market needs. These research directions reflect the imbalance of existing IS research with respect to methodological orientation, which so far focuses on designing novel XAI artifacts and rather neglects to generate empirical insights and develop theory.

Implications

Our findings have implications for different stakeholders of XAI research. IS researchers might benefit from our findings in three different ways. First, the accumulated knowledge helps novice researchers find access to XAI research in IS and assists more experienced researchers in situating their own work in the academic discussion. Second, the presented state of knowledge as well as the future research agenda can inspire researchers to identify research themes that might be of interest to future work. Third, our findings on XAI-receptive publication outlets may assist researchers in identifying potential outlets for their work. Furthermore, editors and reviewers are supported in assessing whether the research under review has sufficiently referenced the existing body of knowledge on XAI in IS and to what extent articles under review are innovative in this field. Finally, given that IS research predominantly addresses business needs (Hevner et al., 2004), our findings are particularly suitable for helping practitioners to make use of the accumulated knowledge on XAI.

Limitations

The findings of this paper have to be seen in light of some limitations. Although we conducted a broad and structured literature search, there exists the possibility that not all relevant articles were identified, due to three reasons. First, while we covered all major IS journals and conferences, the number of sources selected for our literature search is nevertheless limited. Second, although we thoroughly deducted the search terms based on existing XAI literature, additional terms might have revealed further relevant papers. We tried to mitigate this issue by conducting a forward and backward search. Third, by focusing on opaque AI systems, we excluded papers that deal with the explainability of inherently transparent systems, such as rule-based expert systems. Apart from this, by utilizing a quantitative clustering approach to identify research areas, our results do not represent the only possible solution to synthesize existing IS knowledge on XAI. However, our methodology yields a broad, transparent, and replicable overview of XAI research in IS. We hope our findings will help researchers and practitioners gain a thorough overview and better understanding of the body of IS literature on XAI and stimulate further research in this fascinating field.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12525-023-00644-5.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The data that support the findings of this study are available from the authors upon reasonable request.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–18). http://dl.acm.org/citation.cfm?doid=3173574.3174156

Abdul, A., Weth, C. von der, Kankanhalli, M., & Lim, B. Y. (2020). COGAM: Measuring and moderating cognitive load in machine learning model explanations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–14). https://doi.org/10.1145/3313831.3376615

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

Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets, 31*(2), 427–445. https://doi.org/10.1007/s12525-020-00414-7

Aghaeipoor, F., Javidi, M. M., & Fernandez, A. (2021). IFC-BD: An interpretable fuzzy classifier for boosting explainable artificial intelligence in big data. *IEEE Transactions on Fuzzy Systems.* Advance online publication.https://doi.org/10.1109/TFUZZ.2021.3049911

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management, 60*, 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387

Akter, S., Hossain, M. A., Lu, Q. S., & Shams, S. R. (2021b). Big data-driven strategic orientation in international marketing. *International Marketing Review, 38*(5), 927–947. https://doi.org/10.1108/IMR-11-2020-0256

Alam, L., & Mueller, S. (2021). Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics and Decision Making, 21*(1), 1–15. https://doi.org/10.1186/s12911-021-01542-6

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*(1), 1–9. https://doi.org/10.1186/s12911-020-01332-6

Areosa, I., & Torgo, L. (2019). Visual interpretation of regression error. In P. Moura Oliveira, P. Novais, & L. P. Reis (Eds.), *Lecture notes in computer science. Progress in artificial intelligence* (pp. 473–485). Springer International Publishing. https://doi.org/10.1007/978-3-030-30244-3_39

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

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T. & Salovaara, A. (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems, 22*(2). https://aisel.aisnet.org/jais/vol22/iss2/8

Australian Broadcasting Corporation. (2022). *Robodebt inquiry: Royal commission on unlawful debt scheme begins.* ABC News. https://www.youtube.com/results?search_query=robodebt+royal+commission. Accessed 02 Feb 2023

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*, *45*(1). https://doi.org/10.25300/MISQ/2021/15882

Balijepally, V., Mangalaraj, G., & Iyengar, K. (2011). Are we wielding this hammer correctly? A reflective review of the application of cluster analysis in information systems research. *Journal of the Association for Information Systems, 12*(5), 375–413. https://doi.org/10.17705/1jais.00266

Bandara, W., Miskon, S., & Fielt, E. (2011). A systematic, tool-supported method for conducting literature reviews in information systems. *Proceedings of the 19th European Conference on Information Systems (ECIS 2011)* (p. 221). Helsinki, Finland. https://eprints.qut.edu.au/42184/1/42184c.pdf

Barakat, N. H., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Transactions on Information Technology in Biomedicine, 14*(4), 1114–1120. https://doi.org/10.1109/TITB.2009.2039485

Barda, A. J., Horvat, C. M., & Hochheiser, H. (2020). A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Medical Informatics and Decision Making, 20*(1), 1–16. https://doi.org/10.1186/s12911-020-01276-x

Barrera Ferro, D., Brailsford, S., Bravo, C., & Smith, H. (2020). Improving healthcare access management by predicting patient no-show behaviour. *Decision Support Systems*, *138*(113398). https://doi.org/10.1016/j.dss.2020.113398

Bastos, J. A., & Matos, S. M. (2021). Explainable models of credit losses. *European Journal of Operational Research, 301*(1), 386–394. https://doi.org/10.1016/j.ejor.2021.11.009

Bauer, I., Zavolokina, L., & Schwabe, G. (2020). Is there a market for trusted car data? *Electronic Markets, 30*(2), 211–225. https://doi.org/10.1007/s12525-019-00368-5

Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Expl(AI)n it to me – Explainable AI and information systems research. *Business & Information Systems Engineering, 63*, 79–82. https://doi.org/10.1007/s12599-021-00683-2

Bayer, S., Gimpel, H., & Markgraf, M. (2021). The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*, 1–29. https://doi.org/10.1080/12460125.2021.1958505

Beese, J., Haki, M. K., Aier, S., & Winter, R. (2019). Simulation-based research in information systems. *Business & Information Systems Engineering, 61*(4), 503–521. https://doi.org/10.1007/s12599-018-0529-1

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

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect XAI-assisted decision-making: A systematic review. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 78–91). https://hal.telecom-paris.fr/hal-03684457

Blanco-Justicia, A., Domingo-Ferrer, J., Martinez, S., & Sanchez, D. (2020). Machine learning explainability via microaggregation and shallow decision trees. *Knowledge-Based Systems, 194*(5), 105532. https://doi.org/10.1016/j.knosys.2020.105532

Bohanec, M., Kljajić Borštnar, M., & Robnik-Šikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications, 71*(0957–4174), 416–428. https://doi.org/10.1016/j.eswa.2016.11.010

Bresso, E., Monnin, P., Bousquet, C., Calvier, F.-E., Ndiaye, N.-C., Petitpain, N., Smaïl-Tabbone, M., & Coulet, A. (2021). Investigating ADR mechanisms with explainable AI: A feasibility study with knowledge graph mining. *BMC Medical Informatics and Decision Making, 21*(1), 1–14. https://doi.org/10.1186/s12911-021-01518-6

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). *Notes from the AI frontier: Modeling the impact of AI on the world economy*. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

Bunde, E. (2021). AI-assisted and explainable hate speech detection for social media moderators – A design science approach. *Proceedings of the 2021 Annual Hawaii International Conference on System Sciences (HICSS)* (pp. 1264–1274). https://doi.org/10.24251/HICSS.2021.154

Burdisso, S. G., Errecalde, M., & Montes-y-Gómez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications, 133*, 182–197. https://doi.org/10.1016/j.eswa.2019.05.023

Burkart, N., Robert, S., & Huber, M. F. (2021). Are you sure? Prediction revision in automated decision-making. *Expert Systems, 38*(1), e12577. https://doi.org/10.1111/exsy.12577

Chakraborty, D., Başağaoğlu, H., & Winterle, J. (2021). Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Systems with Applications*, *170*(114498). https://doi.org/10.1016/j.eswa.2020.114498

Chakrobartty, S., & El-Gayar, O. (2021). Explainable artificial intelligence in the medical domain: a systematic review. *AMCIS 2021 Proceedings* (p. 1). https://scholar.dsu.edu/cgi/viewcontent.cgi?article=1265&context=bispapers

Chatzimparmpas, A., Martins, R. M., & Kerren, A. (2020). T-viSNE: Interactive assessment and interpretation of t-SNE projections. *IEEE Transactions on Visualization and Computer Graphics, 26*(8), 2696–2714. https://doi.org/10.1109/TVCG.2020.2986996

Cheng, F., Ming, Y., & Qu, H. (2021). Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics, 27*(2), 1438–1447. https://doi.org/10.1109/TVCG.2020.3030342

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–12). https://doi.org/10.1145/3290605.3300789

Chromik, M., & Butz, A. (2021). Human-XAI interaction: A review and design principles for explanation user interfaces. *2021 IFIP Conference on Human-Computer Interaction (INTERACT)* (pp. 619–640). https://doi.org/10.1007/978-3-030-85616-8_36

Chromik, M., & Schuessler, M. (2020). A taxonomy for human subject evaluation of black-box explanations in XAI. *Proceedings of the IUI workshop on explainable smart systems and algorithmic transparency in emerging technologies (ExSS-ATEC'20)* (p. 7). Cagliari, Italy. https://ceur-ws.org/Vol-2582/paper9.pdf

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management, 60*, 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence, 298*, 1–23. https://doi.org/10.1016/j.artint.2021.103503

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society, 1*(1), 104–126. https://doi.org/10.1007/BF03177550

Cooper, A. (2004). *The inmates are running the asylum. Why high-tech products drive us crazy and how to restore the sanity* (2nd ed.). Sams Publishing.

Cui, X., Lee, J. M., & Hsieh, J. P. A. (2019). An integrative 3C evaluation framework for explainable artificial intelligence. *Proceedings of the twenty-fifth Americas conference on information systems (AMCIS),* Cancun, 2019. https://aisel.aisnet.org/amcis2019/ai_semantic_for_intelligent_info_systems/ai_semantic_for_intelligent_info_systems/10

DARPA. (2018). *Explainable artificial intelligence*. https://www.darpa.mil/program/explainable-artificial-intelligence. Accessed 02 Feb 2023

de Bruijn, H., Warnier, M., & Janssen, M. (2021). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly, 39*(2), 101666. https://doi.org/10.1016/j.giq.2021.101666

de Santana, Á. L., Francês, C. R., Rocha, C. A., Carvalho, S. V., Vijaykumar, N. L., Rego, L. P., & Costa, J. C. (2007). Strategies for improving the modeling and interpretability of Bayesian networks. *Data & Knowledge Engineering, 63*, 91–107. https://doi.org/10.1016/j.datak.2006.10.005

Dodge, J., Penney, S., Hilderbrand, C., Anderson, A., & Burnett, M. (2018). How the experts do it: Assessing and explaining agent behaviors in real-time strategy games. *Proceedings of the 36th International Conference on Human Factors in Computing Systems (CHI)* (pp. 1–12). Association for Computing. https://doi.org/10.1145/3173574.3174136

Doran, D., Schulz, S., & Besold, T. R. (2018). What does explainable AI really mean? A new conceptualization of perspectives. In T. R. Besold & O. Kutz (Chairs), *Proceedings of the first international workshop on comprehensibility and explanation in AI and ML 2017*. https://ceur-ws.org/Vol-2071/CExAIIA_2017_paper_2.pdf

Doshi-Velez, F., & Kim, B. (2018). Considerations for evaluation and generalization in interpretable machine learning. In *Explainable and Interpretable Models in Computer Vision and Machine Learning* (pp. 3–17). Springer. https://doi.org/10.1007/978-3-319-98131-4_1

Eiras-Franco, C., Guijarro-Berdiñas, B., Alonso-Betanzos, A., & Bahamonde, A. (2019). A scalable decision-tree-based method to explain interactions in dyadic data. *Decision Support Systems*, *127*(113141). https://doi.org/10.1016/j.dss.2019.113141

Elshawi, R., Al-Mallah, M. H., & Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, *19*(146). https://doi.org/10.1186/s12911-019-0874-0

European Commission (Ed.). (2021). *Regulation of the European Parliament and of the Council: Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN. Accessed 02 Feb 2023

Fang, H. S. A., Tan, N. C., Tan, W. Y., Oei, R. W., Lee, M. L., & Hsu, W. (2021). Patient similarity analytics for explainable clinical risk prediction. *BMC Medical Informatics and Decision Making, 21*(1), 1–12. https://doi.org/10.1186/s12911-021-01566-y

Fernandez, C., Provost, F., & Han, X. (2019). Counterfactual explanations for data-driven decisions. *Proceedings of the fortieth international conference on information systems (ICIS).* https://aisel.aisnet.org/icis2019/data_science/data_science/8

Ferreira, J. J., & Monteiro, M. S. (2020). What are people doing about XAI user experience? A survey on AI explainability research and practice. *2020 International Conference on Human-Computer*

Interaction (HCII) (pp. 56–73). https://doi.org/10.1007/978-3-030-49760-6_4

Fleiß, J., Bäck, E., & Thalmann, S. (2020). Explainability and the intention to use AI-based conversational agents. An empirical investigation for the case of recruiting. *CEUR Workshop Proceedings (CEUR-WS.Org)* (vol 2796, pp. 1–5). https://ceur-ws.org/Vol-2796/xi-ml-2020_fleiss.pdf

Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. *Expert Systems with Applications, 42*(13), 5737–5753. https://doi.org/10.1016/j.eswa.2015.02.042

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020a). Evaluating explainable artificial intelligence – what users really appreciate. *Proceedings of the 2020 European Conference on Information Systems (ECIS). A Virtual AIS Conference.* https://web.archive.org/web/20220803134652id_/https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1194&context=ecis2020_rp

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020b). Fostering human agency: a process for the design of user-centric XAI systems. *In Proceedings of the Forty-First International Conference on Information Systems (ICIS). A Virtual AIS Conference*. https://aisel.aisnet.org/icis2020/hci_artintel/hci_artintel/12

Förster, M., Hühn, P., Klier, M., & Kluge, K. (2021). Capturing users' reality: a novel approach to generate coherent counterfactual explanations. *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS). A Virtual AIS Conference*. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/947e7f6b-c7b0-4dba-afcc-95c4edef0a27/content

Ganeshkumar, M., Ravi, V., Sowmya, V., Gopalakrishnan, E. A., & Soman, K. P. (2021). Explainable deep learning-based approach for multilabel classification of electrocardiogram. *IEEE Transactions on Engineering Management*, 1–13. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9537612&casa_token=6VeV8vXBRT0AAAAA:cVhYpdlNbD1BgRH_9GBDQofEVy38quzW6zs3v3doJzJ2Fx2MP02wy0YqLcoAeC8y2GekDshY0bg&tag=1

Gerlings, J., Shollo, A., & Constantiou, I. (2021). Reviewing the need for explainable artificial intelligence (XAI). *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)* (pp. 1284–1293). https://doi.org/10.48550/arXiv.2012.01007

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine, 178*(11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 80–89). https://doi.org/10.48550/arXiv.1806.00069

Giudici, P., & Raffinetti, E. (2021). Shapley-Lorenz eXplainable Artificial Intelligence. *Expert Systems with Applications*, *167*(114104). https://doi.org/10.1016/j.eswa.2020.114104

Gonzalez, G. (2018). *How Amazon accidentally invented a sexist hiring algorithm: A company experiment to use artificial intelligence in hiring inadvertently favored male candidates.* https://www.inc.com/guadalupe-gonzalez/amazon-artificial-intelligence-ai-hiring-tool-hr.html

Google (Ed.). (2022). *Explainable AI*. https://cloud.google.com/explainable-ai. Accessed 02 Feb 2023

Granados, N., Gupta, A., & Kauffman, R. J. (2010). Information transparency in business-to-consumer markets: Concepts, framework, and research agenda. *Information Systems Research, 21*(2), 207–226. https://doi.org/10.1287/isre.1090.0249

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly, 23*(4), 497–530. https://doi.org/10.2307/249487

Grisci, B. I., Krause, M. J., & Dorn, M. (2021). Relevance aggregation for neural networks interpretability and knowledge discovery on tabular data. *Information Sciences, 559*, 111–129. https://doi.org/10.1016/j.ins.2021.01.052

Gronau, I., & Moran, S. (2007). Optimal implementations of UPGMA and other common clustering algorithms. *Information Processing Letters, 104*(6), 205–210. https://doi.org/10.1016/j.ipl.2007.07.002

Gu, D., Li, Y., Jiang, F., Wen, Z., Liu, S., Shi, W., Lu, G., & Zhou, C. (2020). VINet: A visually interpretable image diagnosis network. *IEEE Transactions on Multimedia, 22*(7), 1720–1729. https://doi.org/10.1109/TMM.2020.2971170

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42. https://doi.org/10.1145/3236009

Guo, M., Xu, Z., Zhang, Q., Liao, X., & Liu, J. (2021). Deciphering feature effects on decision-making in ordinal regression problems: An explainable ordinal factorization model. *ACM Transactions on Knowledge Discovery from Data (TKDD), 16*(3), 1–26. https://doi.org/10.1145/3487048

Ha, T., Sah, Y. J., Park, Y., & Lee, S. (2022). Examining the effects of power status of an explainable artificial intelligence system on users' perceptions. *Behaviour & Information Technology, 41*(5), 946–958. https://doi.org/10.1080/0144929X.2020.1846789

Hamm, P., Wittmann, H. F., & Klesel, M. (2021). Explain it to me and I will use it: A proposal on the impact of explainable AI on use behavior. *ICIS 2021 Proceedings, 9*, 1–9.

Hardt, M., Chen, X., Cheng, X., Donini, M., Gelman, J., Gollaprolu, S., He, J., Larroy, P., Liu, X., McCarthy, N., Rathi, A., Rees, S., Siva, A., Tsai, E., Vasist, K., Yilmaz, P., Zafar, M. B., Das, S., Haas, K., Hill, T., Kenthapadi, K. (2021). Amazon SageMaker clarify: machine learning bias detection and explainability in the cloud. In *2021 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 2974–2983). https://arxiv.org/pdf/2109.03285.pdf

Hatwell, J., Gaber, M. M., & Atif Azad, R. M. (2020). Ada-WHIPS: Explaining AdaBoost classification with applications in the health sciences. *BMC Medical Informatics and Decision Making, 20*(250), 1–25. https://doi.org/10.1186/s12911-020-01201-2

He, J., Hu, H.-J., Harrison, R., Tai, P. C., & Pan, Y. (2006). Transmembrane segments prediction and understanding using support vector machine and decision tree. *Expert Systems with Applications, 30*, 64–72. https://doi.org/10.1016/j.eswa.2005.09.045

Hepenstal, S., Zhang, L., Kodagoda, N., Wong, B., & l. w. (2021). Developing conversational agents for use in criminal investigations. *ACM Transactions on Interactive Intelligent Systems (TiiS), 11*(3–4), 1–35. https://doi.org/10.1145/3444369

Herse, S., Vitale, J., Tonkin, M., Ebrahimian, D., Ojha, S., Johnston, B., Judge, W., & Williams, M. (2018). Do you trust me, blindly? Factors influencing trust towards a robot recommender system. *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. https://ieeexplore.ieee.org/document/8525581/

Heuillet, A., Couthouis, F., & Díaz-Rodríguez, N. (2021). Explainability in deep reinforcement learning. *Knowledge-Based Systems, 214*, 106685. https://doi.org/10.1016/j.knosys.2020.106685

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75–105. https://doi.org/10.2307/25148625

Hong, S. R., Hullman, J., & Bertini, E. (2020). Human factors in model interpretability: Industry practices, challenges, and

needs. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1, Article 68). https://doi.org/10.1145/3392878

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems, 51*(1), 141–154. https://doi.org/10.1016/j.dss.2010.12.003

Iadarola, G., Martinelli, F., Mercaldo, F., & Santone, A. (2021). Towards an interpretable deep learning model for mobile malware detection and family identification. *Computers & Security, 105*, 1–15. https://doi.org/10.1016/j.cose.2021.102198

IBM (Ed.). (2022). *IBM Watson OpenScale - Overview*. https://www.ibm.com/docs/en/cloud-paks/cp-data/3.5.0?topic=services-watson-openscale

Irarrázaval, M. E., Maldonado, S., Pérez, J., & Vairetti, C. (2021). Telecom traffic pumping analytics via explainable data science. *Decision Support Systems, 150*, 1–14. https://doi.org/10.1016/j.dss.2021.113559

Islam, M. A., Anderson, D. T., Pinar, A., Havens, T. C., Scott, G., & Keller, J. M. (2020). Enabling explainable fusion in deep learning with fuzzy integral neural networks. *IEEE Transactions on Fuzzy Systems, 28*(7), 1291–1300. https://doi.org/10.1109/TFUZZ.2019.2917124

Jakulin, A., Možina, M., Demšar, J., Bratko, I., & Zupan, B. (2005). Nomograms for visualizing support vector machines. *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD)* (pp. 108–117). https://doi.org/10.1145/1081870.1081886

Jiang, J., & Cameron, A.-F. (2020). IT-enabled self-monitoring for chronic disease self-management: An interdisciplinary review. *MIS Quarterly, 44*(1), 451–508. https://doi.org/10.25300/MISQ/2020/15108

Jiang, J., Karran, A. J., Coursaris, C. K., Léger, P. M., & Beringer, J. (2022). A situation awareness perspective on human-AI interaction: Tensions and opportunities. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2022.2093863

Jussupow, E., Meza Martínez, M. A., Mädche, A., & Heinzl, A. (2021). Is this system biased? – How users react to gender bias in an explainable AI System. *Proceedings of the 42nd International Conference on Information Systems (ICIS)* (pp. 1–17). https://aisel.aisnet.org/icis2021/hci_robot/hci_robot/11

Kim, C., Lin, X., Collins, C., Taylor, G. W., & Amer, M. R. (2021). Learn, generate, rank, explain: A case study of visual explanation by generative machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS), 11*(3–4), 1–34.

Kim, B., Park, J., & Suh, J. (2020a). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, *134*(113302). https://doi.org/10.1016/j.dss.2020.113302

Kim, J., Lee, S., Hwang, E., Ryu, K. S., Jeong, H., Lee, J. W., Hwangbo, Y., Choi, K. S., & Cha, H. S. (2020b). Limitations of deep learning attention mechanisms in clinical research: Empirical case study based on the Korean diabetic disease setting. *Journal of Medical Internet Research*, *22*(12). https://doi.org/10.2196/18418

Kliegr, T., Bahník, Š, & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence, 295*, 103458. https://doi.org/10.1016/j.artint.2021.103458

Kline, A., Kline, T., Shakeri Hossein Abad, Z., & Lee, J. (2020). Using item response theory for explainable machine learning in predicting mortality in the intensive care unit: Case-based approach. *Journal of Medical Internet Research*, *22*(9). https://doi.org/10.2196/20268

Knowles, T. (2021). *AI will have a bigger impact than fire, says Google boss Sundar Pichai*. https://www.thetimes.co.uk/article/ai-will-have-a-bigger-impact-than-fire-says-google-boss-sundar-pichai-rk8bdst7r

Kou, Y., & Gui, X. (2020). Mediating community-AI interaction through situated explanation. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2, Article 102). https://doi.org/10.1145/3415173

Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2020). Generating and understanding personalized explanations in hybrid recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS), 10*(4), 1–40.

Kumar, A., Manikandan, R., Kose, U., Gupta, D., & Satapathy, S. C. (2021). Doctor's dilemma: Evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 17*(3s), 1–26.

Kute, D. V., Pradhan, B., Shukla, N., & Alamri, A. (2021). Deep learning and explainable artificial intelligence techniques applied for detecting money laundering – A critical review. *IEEE Access, 9*, 82300–82317.

Kwon, B. C., Choi, M.-J., Kim, J. T., Choi, E., Kim, Y. B., Kwon, S., Sun, J., & Choo, J. (2019). Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1). https://doi.org/10.1109/TVCG.2018.2865027

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

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Seeing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*. https://doi.org/10.1016/j.artint.2021.103473

Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science*, *9*. https://doi.org/10.28945/479

Li, J., Shi, H., & Hwang, K. S. (2021). An explainable ensemble feedforward method with Gaussian convolutional filter. *Knowledge-Based Systems*, *225*. https://doi.org/10.1016/j.knosys.2021.107103

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 (CHI)* (pp. 1–15) https://doi.org/10.1145/3313831.3376590

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Proceedings of the 2009 SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 2119–2128). https://doi.org/10.1145/1518701.1519023

Lopez-Gazpio, I., Maritxalar, M., Gonzalez-Agirre, A., Rigau, G., Uria, L., & Agirre, E. (2017). Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems, 119*, 186–199. https://doi.org/10.1016/j.knosys.2016.12.013

Lukyanenko, R., Castellanos, A., Storey, V. C., Castillo, A., Tremblay, M. C., & Parsons, J. (2020). Superimposition: Augmenting machine learning outputs with conceptual models for explainable AI. In G. Grossmann & S. Ram (Eds.), *Lecture notes in computer science. Advances in conceptual modeling* (pp. 26–34). Springer International Publishing. https://doi.org/10.1007/978-3-030-65847-2_3

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures, 90*, 46–60. https://doi.org/10.1016/j.futures.2017.03.006

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press.

Marella, V., Upreti, B., Merikivi, J., & Tuunainen, V. K. (2020). Understanding the creation of trust in cryptocurrencies: The case of Bitcoin. *Electronic Markets, 30*(2), 259–271. https://doi.org/10.1007/s12525-019-00392-5

Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. *MIS Quarterly, 38*(1), 73–99. https://doi.org/10.25300/MISQ/2014/38.1.04

Martens, D., Baesens, B., & van Gestel, T. (2009). Decompositional rule extraction from support vector machines by active learning. *IEEE Transactions on Knowledge and Data Engineering, 21*(2), 178–191. https://doi.org/10.1109/TKDE.2008.131

Martens, D., Baesens, B., van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *SSRN Electronic Journal*. Advance online publication.https://doi.org/10.2139/ssrn.878283

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J. & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577* (7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Mehdiyev, N., & Fettke, P. (2020). Prescriptive process analytics with deep learning and explainable artificial intelligence. *Proceedings of the 28th European Conference on Information Systems (ECIS)*. An Online AIS Conference. https://aisel.aisnet.org/ecis2020_rp/122

Mensa, E., Colla, D., Dalmasso, M., Giustini, M., Mamo, C., Pitidis, A., & Radicioni, D. P. (2020). Violence detection explanation via semantic roles embeddings. *BMC Medical Informatics and Decision Making*, *20*(263). https://doi.org/10.1186/s12911-020-01237-4

Merry, M., Riddle, P., & Warren, J. (2021). A mental models approach for defining explainable artificial intelligence. *BMC Medical Informatics and Decision Making, 21*(1), 1–12. https://doi.org/10.1186/s12911-021-01703-7

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2020). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management, 39*(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Meske, C., Abedin, B., Klier, M., & Rabhi, F. (2022). Explainable and responsible artificial intelligence. *Electronic Markets, 32*(4), 2103–2106. https://doi.org/10.1007/s12525-022-00607-2

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

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. *ArXiv*. arXiv:1712.00547. https://arxiv.org/pdf/1712.00547.pdf

Ming, Y., Huamin, Qu., & Bertini, E. (2019). RuleMatrix: Visualizing and understanding classifiers with rules. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 342–352. https://doi.org/10.1109/TVCG.2018.2864812

Mirbabaie, M., Brendel, A. B., & Hofeditz, L. (2022). Ethics and AI in information systems research. *Communications of the Association for Information Systems, 50*(1), 38. https://doi.org/10.17705/1CAIS.05034

Mitra, S., & Hayashi, Y. (2000). Neuro-fuzzy rule generation: Survey in soft computing framework. *IEEE Transactions on Neural Networks, 11*(3), 748–768. https://doi.org/10.1109/72.846746

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

Mombini, H., Tulu, B., Strong, D., Agu, E. O., Lindsay, C., Loretz, L., Pedersen, P., & Dunn, R. (2021). An explainable machine learning model for chronic wound management decisions. *AMCIS 2021 Proceedings, 18*, 1–10.

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal, 73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

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

Moreira, C., Chou, Y.-L., Velmurugan, M., Ouyang, C., Sindhgatta, R., & Bruza, P. (2021). LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models. *Decision Support Systems, 150*, 1–16. https://doi.org/10.1016/j.dss.2021.113561

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications, 165*, 1–8. https://doi.org/10.1016/j.eswa.2020.113986

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *ArXiv*. https://arxiv.org/pdf/1902.01876

Murray, B. J., Islam, M. A., Pinar, A. J., Anderson, D. T., Scott, G. J., Havens, T. C., & Keller, J. M. (2021). Explainable AI for the Choquet integral. *IEEE Transactions on Emerging Topics in Computational Intelligence, 5*(4), 520–529. https://doi.org/10.1109/TETCI.2020.3005682

Narayanan, M., Chen, E., He, J, Kim, B, Gershman, S., & Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *ArXiv,* 1802.00682. https://doi.org/10.48550/arXiv.1802.00682

Nascita, A., Montieri, A., Aceto, G., Ciuonzo, D., Persico, V., & Pescapé, A. (2021). XAI meets mobile traffic classification: Understanding and improving multimodal deep learning architectures. *IEEE Transactions on Network and Service Management, 18*(4), 4225–4246. https://doi.org/10.1109/TNSM.2021.3098157

Neto, M. P., & Paulovich, F. V. (2021). Explainable matrix - visualization for global and local interpretability of random forest classification ensembles. *IEEE Transactions on Visualization and Computer Graphics, 27*(2), 1427–1437. https://doi.org/10.1109/TVCG.2020.3030354

Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction, 27*(3), 393–444. https://doi.org/10.1007/s11257-017-9195-0

Omeiza, D., Webb, H., Jirotka, M., & Kunze, L. (2021). Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems, 23*(8), 10142–10162. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9616449&casa_token=pCkvj82hzqwAAAAA:yYPZ8qTUP7U8tLQj793sviDzuwLewzQZCvBPza4SHtG_P-eSlpp0Te5X9aF1OuVt35wT6EMfP1w&tag=1

Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., Liu, X., & He, Z. (2020). Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. *Journal of the American Medical*

Informatics Association: JAMIA, 27(7), 1173–1185. https://doi.org/10.1093/jamia/ocaa053

Peñafiel, S., Baloian, N., Sanson, H., & Pino, J. A. (2020). Applying Dempster-Shafer theory for developing a flexible, accurate and interpretable classifier. *Expert Systems with Applications, 148*(113262), 1–12. https://doi.org/10.1016/j.eswa.2020.113262

Pessach, D., Singer, G., Avrahami, D., Chalutz Ben-Gal, H., Shmueli, E., & Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, *134*(113290). https://doi.org/10.1016/j.dss.2020.113290

Pierrard, R., Poli, J.-P., & Hudelot, C. (2021). Spatial relation learning for explainable image classification and annotation in critical applications. *Artificial Intelligence*, *292*(103434). https://doi.org/10.1016/j.artint.2020.103434

Probst, F., Grosswiele, L., & Pfleger, R. (2013). Who will lead and who will follow: Identifying Influential Users in Online Social Networks. *Business & Information Systems Engineering, 5*(3), 179–193. https://doi.org/10.1007/s12599-013-0263-7

Rader, E., & Gray, R. (2015). Understanding user beliefs about algorithmic curation in the Facebook news feed. *Proceedings of the 33rd International Conference on Human Factors in Computing Systems (CHI)* (pp. 173–182). https://doi.org/10.1145/2702123.2702174

Ragab, A., El-Koujok, M., Poulin, B., Amazouz, M., & Yacout, S. (2018). Fault diagnosis in industrial chemical processes using interpretable patterns based on Logical Analysis of Data. *Expert Systems with Applications, 95*, 368–383. https://doi.org/10.1016/j.eswa.2017.11.045

Rana, N. P., Chatterjee, S., Dwivedi, Y. K., & Akter, S. (2022). Understanding dark side of artificial intelligence (AI) integrated business analytics: Assessing firm's operational inefficiency and competitiveness. *European Journal of Information Systems, 31*(3), 364–387. https://doi.org/10.1080/0960085X.2021.1955628

Rawal, A., McCoy, J., Rawat, D., Sadler, B., & Amant, R. (2021). Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives. *IEEE Transactions on Artificial Intelligence, 1*(01), 1–1. https://doi.org/10.1109/TAI.2021.3133846

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? A proposal of user-centered explainable AI. In C. Trattner, D. Parra, & N. Riche (Chairs), *Joint Proceedings of the ACM IUI 2019 Workshops*. http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf

Rissler, R., Nadj, M., Adam, M., & Maedche, A. (2017). Towards an integrative theoretical Framework of IT-Mediated Interruptions. *Proceedings of the 25th European Conference on Information Systems (ECIS)*. http://aisel.aisnet.org/ecis2017_rp/125

Robert, L. P., Bansal, G., & Lütge, C. (2020). ICIS 2019 SIGHCI Workshop Panel Report: Human– computer interaction challenges and opportunities for fair, trustworthy and ethical artificial intelligence. *AIS Transactions on Human-Computer Interaction, 12*(2), 96–108. https://doi.org/10.17705/1thci.00130

Rowe, F. (2014). What literature review is not: Diversity, boundaries and recommendations. *European Journal of Information Systems, 23*(3), 241–255. https://doi.org/10.1057/ejis.2014.7

Russell, S., & Norvig, P. (2021). *Artificial intelligenc: A modern approach (4th)*. Pearson.

Rzepka, C., & Berger, B. (2018). User interaction with AI-enabled systems: A systematic review of IS research. *Proceedings of the Thirty-Nine International Conference on Information Systems (ICIS)*. https://aisel.aisnet.org/icis2018/general/Presentations/7

Sachan, S., Yang, J.-B., Xu, D.-L., Benavides, D. E., & Li, Y. (2020). An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications, 144*(113100), 1–49. https://doi.org/10.1016/j.eswa.2019.113100

Schlicker, N., Langer, M., Ötting, S. K., Baum, K., König, C. J., & Wallach, D. (2021). What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior, 122*, 1–16. https://doi.org/10.1016/j.chb.2021.106837

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*. Advance online publication. https://doi.org/10.1080/12460125.2020.1819094

Schneider, J., & Handali, J. P. (2019). Personalized explanation for machine learning: a conceptualization. *Proceedings of the Twenty-Seventh European Conference on Information Systems (ECIS 2019)*. Stockholm-Uppsala, Sweden. https://arxiv.org/ftp/arxiv/papers/1901/1901.00770.pdf

Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. *Expert Systems with Applications, 41*(5), 2239–2249. https://doi.org/10.1016/j.eswa.2013.09.022

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), 233–242. https://doi.org/10.1093/idpl/ipx022

Sevastjanova, R., Jentner, W., Sperrle, F., Kehlbeck, R., Bernard, J., & El-Assady, M. (2021). QuestionComb: A gamification approach for the visual explanation of linguistic phenomena through interactive labeling. *ACM Transactions on Interactive Intelligent Systems (TiiS), 11*(3–4), 1–38.

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 747–748). https://doi.org/10.1109/DSAA49011.2020.00096

Sharma, P., Mirzan, S. R., Bhandari, A., Pimpley, A., Eswaran, A., Srinivasan, S., & Shao, L. (2020). Evaluating tree explanation methods for anomaly reasoning: A case study of SHAP TreeExplainer and TreeInterpreter. In G. Grossmann & S. Ram (Eds.), *Lecture notes in computer science. Advances in conceptual modeling* (pp. 35–45). Springer International Publishing. https://doi.org/10.1007/978-3-030-65847-2_4

Shen, H., Jin, H., Cabrera, Á. A., Perer, A., Zhu, H., & Hong, J. I. (2020). Designing alternative representations of confusion matrices to support non-expert public understanding of algorithm performance. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), 1–22. https://doi.org/10.1145/3415224

Shin, D. (2021a). 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

Shin, D. (2021b). Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm. *Journal of Information Science*, 1–14. https://doi.org/10.1177/0165551520985495

Sidorova, A., Evangelopoulos, N., Valacich, J. S., & Ramakrishnan, T. (2008). Uncovering the intellectual core of the information systems discipline. *MIS Quarterly*, 467–482. https://www.jstor.org/stable/25148852

Singh, N., Singh, P., & Bhagat, D. (2019). A rule extraction approach from support vector machines for diagnosing hypertension among diabetics. *Expert Systems with Applications, 130*, 188–205. https://doi.org/10.1016/j.eswa.2019.04.029

Soares, E., Angelov, P. P., Costa, B., Castro, M. P. G., Nageshrao, S., & Filev, D. (2021). Explaining deep learning models through rule-based approximation and visualization. *IEEE Transactions*

on Fuzzy Systems, 29(8), 2399–2407. https://doi.org/10.1109/TFUZZ.2020.2999776

Spinner, T., Schlegel, U., Schafer, H., & El-Assady, M. (2020). Explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics, 26*(1), 1064–1074. https://doi.org/10.1109/TVCG.2019.2934629

Springer, A., & Whittaker, S. (2020). Progressive disclosure: When, why, and how do users want algorithmic transparency information? *ACM Transactions on Interactive Intelligent Systems (TiiS), 10*(4), 1–32. https://doi.org/10.1145/3374218

Stoean, R., & Stoean, C. (2013). Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. *Expert Systems with Applications, 40*, 2677–2686. https://doi.org/10.1016/j.eswa.2012.11.007

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems, 41*, 647–665. https://doi.org/10.1007/s10115-013-0679-x

Su, G., Lin, B., Luo, W., Yin, J., Deng, S., Gao, H., & Xu, R. (2021). Hypomimia recognition in Parkinson's disease with semantic features. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *17*(3), 1–20. https://doi.org/10.1145/3476778

Sultana, T., & Nemati, H. (2021). Impact of explainable AI and task complexity on human-machine symbiosis. *Proceedings of the Twenty-Seventh Americas Conference on Information Systems (AMCIS)*. https://aisel.aisnet.org/amcis2021/sig_hci/sig_hci/20

Sun, C., Dui, H., & Li, H. (2021). Interpretable time-aware and co-occurrence-aware network for medical prediction. *BMC Medical Informatics and Decision Making, 21*(1), 1–12.

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication, 25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Tabankov, S. S., & Möhlmann, M. (2021). Artificial intelligence for in-flight services: How the Lufthansa group managed explainability and accuracy concerns. *Proceedings of the International Conference on Information Systems (ICIS), 16*, 1–9.

Taha, I. A., & Ghosh, J. (1999). Symbolic interpretation of artificial neural networks. *IEEE Transactions on Knowledge and Data Engineering, 11*(3), 448–463. https://doi.org/10.1109/69.774103

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

van der Waa, J., Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2020). Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*, *144*(102493). https://doi.org/10.1016/j.ijhcs.2020.102493

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: A systematic review. *ArXiv*. https://arxiv.org/pdf/2006.00093

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*(103404). https://doi.org/10.1016/j.artint.2020.103404

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems, 25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

vom Brocke, J., Simons, A., Niehaves, B [Bjoern], Niehaves, B [Bjorn], Reimer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. *ECIS 2009 Proceedings*(161). http://aisel.aisnet.org/ecis2009/161

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. *Proceedings of the 2019 Conference on Human Factors in Computing Systems (CHI)*. http://dl.acm.org/citation.cfm?doid=3290605.3300831

Wanner, J., Heinrich, K., Janiesch, C., & Zschech, P. (2020a). How much AI do you require decision factors for adopting AI technology. *Proceedings of the Forty-First International Conference on Information Systems (ICIS)*. https://aisel.aisnet.org/icis2020/implement_adopt/implement_adopt/10

Wanner, J., Herm, L. V., & Janiesch, C. (2020b). How much is the black box? The value of explainability in machine learning models. *ECIS 2020 Research-in-Progress*. https://aisel.aisnet.org/ecis2020_rip/85

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly, 26*(2), xiii–xxiii.

Xiong, J., Qureshi, S., & Najjar, L. (2014). A cluster analysis of research in information technology for global development: Where to from here? *Proceedings of the SIG GlobDev Seventh Annual Workshop*. https://aisel.aisnet.org/globdev2014/1

Yampolskiy, R. V. (2019). Predicting future AI failures from historic examples. *Foresight, 21*(1), 138–152. https://doi.org/10.1108/FS-04-2018-0034

Yan, A., & Xu, D. (2021). AI for depression treatment: Addressing the paradox of privacy and trust with empathy, accountability, and explainability. *Proceedings of the Fourty-Second International Conference on Information Systems (ICIS)*. https://aisel.aisnet.org/icis2021/is_health/is_health/15/

Yang, Z., Zhang, A., & Sudjianto, A. (2021). Enhancing explainability of neural networks through architecture constraints. *IEEE Transactions on Neural Networks and Learning Systems, 32*(6), 2610–2621. https://doi.org/10.1109/TNNLS.2020.3007259

Yoo, S., & Kang, N. (2021). Explainable artificial intelligence for manufacturing cost estimation and machining feature visualization. *Expert Systems with Applications, 183*, 1–14. https://doi.org/10.1016/j.eswa.2021.115430

Zeltner, D., Schmid, B., Csiszár, G., & Csiszár, O. (2021). Squashing activation unctions in benchmark tests: Towards a more eXplainable Artificial Intelligence using continuous-valued logic. *Knowledge-Based Systems*, *218*. https://doi.org/10.1016/j.knosys.2021.106779

Zhang, Q. S., & Zhu, S. C. (2018). Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering, 19*(1), 27–39. https://doi.org/10.1631/FITEE.1700808

Zhang, K., Liu, X., Liu, F., He, L., Zhang, L., Yang, Y., Li, W., Wang, S., Liu, L., Liu, Z., Wu, X., & Lin, H. (2018). An interpretable and expandable deep learning diagnostic system for multiple ocular diseases: Qualitative study. *Journal of Medical Internet Research, 20*(11), 1–13. https://doi.org/10.2196/11144

Zhang, C. A., Cho, S., & Vasarhelyi, M. (2022). Explainable Artificial Intelligence (XAI) in auditing. *International Journal of Accounting Information Systems, 46*, 100572. https://doi.org/10.1016/j.accinf.2022.100572

Zhao, X., Wu, Y., Lee, D. L., & Cui, W. (2019). Iforest: Interpreting random forests via visual analytics. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 407–416. https://doi.org/10.1109/TVCG.2018.2864475

Zhdanov, D., Bhattacharjee, S., & Bragin, M. (2021). Incorporating FAT and privacy aware AI modeling approaches into business

decision making frameworks. *Decision Support Systems, 155*, 1–12. https://doi.org/10.1016/j.dss.2021.113715

Zhong, Q., Fan, X., Luo, X., & Toni, F. (2019). An explainable multi-attribute decision model based on argumentation. *Expert Systems with Applications, 117*, 42–61. https://doi.org/10.1016/j.eswa.2018.09.038

Zhu, C., Chen, Z., Zhao, R., Wang, J., & Yan, R. (2021). Decoupled feature-temporal CNN: Explaining deep learning-based machine health monitoring. *IEEE Transactions on Instrumentation and Measurement, 70*, 1–13. https://doi.org/10.1109/TIM.2021.3084310

Zytek, A., Liu, D., Vaithianathan, R., & Veeramachaneni, K. (2021). Sibyl: Explaining machine learning models for high-stakes decision making. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–6). https://doi.org/10.1145/3411763.3451743