Behind the Scenes of Emerging Technologies – Opportunities, Challenges, and Solution Approaches Along a Socio-Technical Continuum

Dissertation

to obtain the doctoral degree of Economic Sciences

(Dr. oec.)

Faculty of Business, Economics and Social Sciences

University of Hohenheim

Institute of Marketing & Management

submitted by

Sarah Bayer

from *Ingolstadt*

2021

Date of disputation: November 10, 2021

Supervisor and first reviewer:	Prof. Dr. Henner Gimpel
Second reviewer:	Prof. Dr. Mareike Schoop
Exam chairperson:	Prof. Dr. Verena Hüttl-Maack
Dean of faculty:	Prof. Dr. Karsten Hadwich

Abstract

Digitalization is a socio-technical phenomenon that shapes our lives as individuals, economies, and societies. The perceived complexity of technologies continues to increase, and technology convergence makes a clear separation between technologies impossible. A good example of this is the Internet of Things (IoT) with its embedded Artificial Intelligence (AI). Furthermore, a separation of the social and the technical component has become near enough impossible, for which there is increasing awareness in the Information Systems (IS) community. Overall, emerging technologies such as AI or IoT are becoming less understandable and transparent, which is evident for instance when AI is described in terms of a “black box”. This opacity undermines humans’ trust in emerging technologies, which, however, is crucial for both its usage and spread, especially as emerging technologies start to perform tasks that bear high risks for humans, such as autonomous driving. Critical perspectives on emerging technologies are often discussed in terms of ethics, including such aspects as the responsibility for decisions made by algorithms, the limited data privacy, and the moral values that are encoded in technology. In sum, the varied opportunities that come with digitalization are accompanied by significant challenges.

Research on the negative ramifications of AI is crucial if we are to foster a human-centered technological development that is not simply driven by opportunities but by utility for humanity. As the IS community is positioned at the intersection of the technological and the social context, it plays a central role in finding answers to the question as to how the advantages outweigh the challenges that come with emerging technologies. Challenges are examined under the label of “dark side of IS”, a research area which receives considerably less attention in existing literature than the positive aspects (Gimpel & Schmied, 2019)[1]. With its focus on challenges, this dissertation aims to counterbalance this. Since the remit of IS research is the entire information system, rather than merely the technology, humanistic and instrumental goals ought to be considered in equal measure. This dissertation follows calls for research for a healthy distribution along the so-called socio-technical continuum (Sarker et al., 2019)[2], that broadens its focus to include the social as well as the technical, rather than looking at one or the other. With that in mind, this dissertation aims to advance knowledge on IS with regard to opportunities, and in particular with a focus on challenges of two emerging technologies, IoT and AI, along the socio-technical continuum.

[1] Gimpel, H., & Schmied, F. (2019). Risks and side effects of digitalization: A multi-level taxonomy of the adverse effects of using digital technologies and media. Proceedings of the 27th European Conference on Information Systems (ECIS2019), Stockholm-Uppsala, Sweden.

[2] Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the IS discipline: Its historical legacy and its continued relevance. Management Information Systems Quarterly, 43(3), 695–719.

This dissertation provides novel insights for individuals to better understand opportunities, but in particular possible negative side effects. It guides organizations on how to address these challenges and suggests not only the necessity of further research along the socio-technical continuum but also several ideas on where to take this future research.

Chapter 2 contributes to research on opportunities and challenges of IoT. Section 2.1 identifies and structures opportunities that IoT devices provide for retail commerce customers. By conducting a structured literature review, affordances are identified, and by examining a sample of 337 IoT devices, completeness and parsimony are validated. Section 2.2 takes a close look at the ethical challenges posed by IoT, also known as IoT ethics. Based on a structured literature review, it first identifies and structures IoT ethics, then provides detailed guidance for further research in this important and yet under-appreciated field of study. Together, these two research articles underline that IoT has the potential to radically transform our lives, but they also illustrate the urgent need for further research on possible ethical issues that are associated with IoTs' specific features.

Chapter 3 contributes to research on AI along the socio-technical continuum. Section 3.1 examines algorithms underlying AI. Through a structured literature review and semi-structured interviews analyzed with a qualitative content analysis, this section identifies, structures and communicates concerns about algorithmic decision-making and is supposed to improve offers and services. Section 3.2 takes a deep dive into the concept of moral agency in AI to discuss whether responsibility in human-computer interaction can be grasped better with the concept of "agency". In section 3.3, data from an online experiment with a self-developed AI system is used to examine the role of a user's domain-specific expertise in trusting and following suggestions from AI decision support systems. Finally, section 3.4 draws on design science research to present a framework for ethical software development that considers ethical issues from the beginning of the design and development process. By looking at the multiple facets of this topic, these four research articles ought to guide practitioners in deciding which challenges to consider during product development. With a view to subsequent steps, they also offer first ideas on how these challenges could be addressed. Furthermore, the articles offer a basis for further, solution-oriented research on AI's challenges and encourage users to form their own, informed, opinions.

In sum, this dissertation contributes to scientific knowledge development in IS research on opportunities, but in particular on challenges that arise from IoT and AI along the socio-technical continuum. The research articles included in this dissertation provide insights for IoTs affordances, examines challenges, in particular ethical challenges, of AI and IoT, and provides ideas for potential solution approaches for the concept of trust in AI and ethics in software development. This dissertation hopefully provides both theoretical and practical contributions for further research on IoT and AI along the socio-technical continuum, research that is driven first and foremost by human needs.

Zusammenfassung

Die Digitalisierung ist ein sozio-technisches Phänomen, das unser persönliches Leben, aber auch die Wirtschaft und die gesamte Gesellschaft prägt. Die wahrgenommene Komplexität von Technologie nimmt stetig zu. Die Technologiekonvergenz macht eine klare Trennung zwischen Technologien praktisch unmöglich, wofür das Internet der Dinge (IoT) mit seiner eingebetteten Künstlichen Intelligenz (KI) ein gutes Beispiel ist. Darüber hinaus wird eine Trennung der sozialen und der technischen Komponente nahezu unmöglich, wofür es ein steigendes Bewusstsein in der Information Systems (IS) Community gibt. Insgesamt werden aufstrebende Technologien wie KI oder IoT weniger verständlich und transparent, was sich beispielsweise darin zeigt, dass KI der Begriff der „Black Box" zugeschrieben wird. Die Undurchsichtigkeit untergräbt das Vertrauen der Menschen in aufstrebende Technologien, das jedoch für die Nutzung und Verbreitung dieser entscheidend ist, insbesondere wenn Technologien Aufgaben übernehmen oder unterstützen, die hohe Risiken für den Menschen bergen, wie z. B. autonomes Fahren. Kritische Perspektiven auf neue Technologien werden oft unter dem Begriff der Ethik diskutiert, darunter Aspekte wie die Verantwortung für Entscheidungen, die von Algorithmen getroffen werden, moralische Werte, die in die Technologie eingebettet sind, und Datenschutz. Zusammenfassend lässt sich sagen, dass die vielfältigen Chancen der Digitalisierung mit Herausforderungen einhergehen.

Die Forschung zu Risiken und Nebenwirkungen ist entscheidend, um eine menschenzentrierte technologische Entwicklung zu fördern, die nicht nur von den Möglichkeiten, sondern insbesondere vom Nutzenstiften für die Menschheit getrieben ist. An der Schnittstelle zwischen Technologie und sozialem Kontext angesiedelt, spielt die IS-Community eine wichtige Rolle bei der Suche nach Antworten auf die Frage, wie die Vorteile die Risiken neuer Technologien überwiegen können. Herausforderungen werden im Forschungsbereich „dark side of IS" untersucht, welcher in der bestehenden Literatur deutlich weniger Aufmerksamkeit erhält als die positiven Aspekte (Gimpel & Schmied, 2019)[3]. Dem möchte diese Dissertation ein Stück weit entgegenwirken, indem ein Fokus auf die Herausforderungen gelegt wird. Da in der IS-Forschung das gesamte Informationssystem und nicht nur die Technologie im Mittelpunkt der Betrachtung steht, sollen humanistische und instrumentelle Ziele gleichermaßen berücksichtigt werden. Darüber hinaus folgt diese Dissertation dem Aufruf nach einer angemessenen Verteilung der

[3] Gimpel, H., & Schmied, F. (2019). Risks and side effects of digitalization: A multi-level taxonomy of the adverse effects of using digital technologies and media. Proceedings of the 27th European Conference on Information Systems (ECIS2019), Stockholm-Uppsala, Sweden.

Forschung entlang des sogenannten sozio-technischen Kontinuums (Sarker et al., 2019)[4] und löst sich somit von Forschung, die am sozialen oder technischen Endpunkt des Kontinuums angesiedelt ist. Zusammenfassend zielt diese Dissertation darauf ab, das Wissen über IS im Hinblick auf die Chancen und insbesondere die Herausforderungen entlang des sozio-technischen Kontinuums der aufkommenden Technologien IoT und KI voranzutreiben. Damit liefert die Dissertation neue Einblicke für Individuen, um die Möglichkeiten, aber insbesondere die potenziellen negativen Nebenwirkungen der Digitalisierung besser zu verstehen, bietet Orientierung für Organisationen, um diese Herausforderungen zu adressieren, und veranschaulicht die Notwendigkeit und Ideen für weitere Forschung entlang des sozio-technischen Kontinuums.

Kapitel 2 leistet einen Beitrag zur Forschung über Chancen und Herausforderungen des IoT. Kapitel 2.1 identifiziert und strukturiert Chancen von IoT-Geräten für Kunden im Einzelhandel. Mit einer strukturierten Literaturrecherche werden Affordanzen von IoT-Geräten für Kunden identifiziert und mit einer Stichprobe von 337 IoT-Geräten wird eine Validierung hinsichtlich Vollständigkeit und Sparsamkeit durchgeführt. Kapitel 2.2 beschäftigt sich mit ethischen Herausforderungen des IoT, genannt IoT-Ethik. Basierend auf einer strukturierten Literaturrecherche identifiziert und strukturiert es die IoT-Ethik und gibt detaillierte Hinweise für die weitere Erforschung dieses wichtigen, aber noch zu wenig erforschten Feldes. Mit diesen beiden Forschungsartikeln unterstreicht diese Dissertation das Potenzial des IoT, unser Leben radikal zu verändern, verdeutlicht aber auch den Bedarf an weiterer Forschung zu potenziellen ethischen Fragen, die mit den spezifischen Eigenschaften des IoT verbunden sind. Kapitel 3 trägt zur Forschung über KI entlang des sozio-technischen Kontinuums bei. Kapitel 3.1 untersucht die Algorithmen, die KI zugrunde liegen. Eine strukturierte Literaturrecherche und semi-strukturierte Interviews, die mit einer qualitativen Inhaltsanalyse analysiert werden, zielen darauf ab, Bedenken gegenüber algorithmischer Entscheidungsfindung zu identifizieren, zu strukturieren und zu kommunizieren, um darauf basierend Angebote und Dienstleistungen zu verbessern. Kapitel 3.2 bietet eine ethische Vertiefung in das Konzept der moralischen Handlungsfähigkeit und untersucht, ob Verantwortung in der Mensch-Computer-Interaktion mit dem Konzept der „Agency“ besser erfasst werden kann. In Kapitel 3.3 wird anhand von Daten aus einem Online-Experiment mit einem selbst entwickelten KI-System untersucht, welche Rolle das domänenspezifische Fachwissen der Nutzer für das Vertrauen in und das Befolgen von Vorschlägen von KI-Entscheidungsunterstützungssystemen spielt. Schließlich wird in Kapitel 3.4

[4] Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the is discipline: Its historical legacy and its continued relevance. Management Information Systems Quarterly, 43(3), 695–719.

auf der Grundlage designwissenschaftlicher Forschung ein Rahmenwerk für ethische Softwareentwicklung vorgestellt, das ethische Aspekte bereits zu Beginn des Design- und Entwicklungsprozesses berücksichtigt. Diese vier Forschungsartikel können Praktikern als Orientierung dienen, welche Herausforderungen bei der Produktentwicklung zu berücksichtigen sind und bieten erste Ideen, wie sie diese angehen können. Darüber hinaus bieten die Forschungsergebnisse eine Grundlage für weitere, lösungsorientierte Forschung zu den Herausforderungen von KI und ermutigen Nutzer, sich eine eigene, fundierte Meinung zu bilden.

Zusammenfassend liefert diese Dissertation wissenschaftliche Erkenntnisse für die IS-Forschung zu Chancen, aber insbesondere zu Herausforderungen von IoT und KI entlang des sozio-technischen Kontinuums. Die in dieser Dissertation enthaltenen Forschungsbeiträge geben Einblicke in die Affordanzen von IoT, untersuchen die Herausforderungen, insbesondere die ethischen Herausforderungen, von KI und IoT und liefern Ideen für mögliche Lösungsansätze für das Konzept des Vertrauens in KI und der Ethik in der Softwareentwicklung. Diese Dissertation soll theoretische und praktische Beiträge für weitere, an den menschlichen Bedürfnissen orientierte Forschung zu IoT und KI entlang des sozio-technischen Kontinuums liefern.

Table of contents

List of tables

List of figures

1 Introduction

1.1 Motivation[1]

> *"Everyone sees the world from their own perspective. We tell ourselves stories about how things came to be and what needs to be done. Such narratives shape the way we see and change the world. And in order to use technology to create a more humane world, we need to take the narratives into our own hands."*
>
> *Bengiamin Barblan (2018)*

For technologies that serves humanity, a multitude of narratives should be included in the software design and development process, which is not limited to software developers, but includes every human that is potentially affected by technology. The dimension of the digitization makes it necessary to consider numerous perspectives in its design and development in order to strive for technology that considers instrumental as well as humanistic outcomes (Sarker et al., 2019). Digitalization is developing at a considerable pace, shaping our lives as individuals, economies, and societies (Berger et al., 2018; Gimpel & Röglinger, 2015; Matt et al., 2019). Associated with digitalization are various opportunities to invent new products, services, and business models, such as a high degree of individualization shaping customer experience (Rachinger et al., 2019). These opportunities are accompanied by challenges such as increasing technological complexity and changing legal requirements (Rachinger et al., 2019). Defined as "manifold socio-technical phenomena and processes of adopting and using [digital] technologies in broader individual, organizational, and societal contexts" (Legner et al., 2017, p. 301), digitalization combines two components, the technical and the social.

There is increasing awareness in the Information Systems (IS) community that a clear separation between the technical and the social component is neither helpful nor indeed possible, as any examination of human interaction with the technical requires a holistic perspective on the socio-technical system (Bednar & Welch, 2020). Due to technology convergence, even a separation between technologies becomes complex to the point of impossibility. The phenomenon of technology convergence has drawn ever greater attention since the 2000s. It describes unclear boundaries of initially separate technologies that are

[1] Since it is in the nature of a cumulative dissertation that it consists of individual research papers, this Chapter (chapter 1) as well as the last chapter (chapter 4) partly comprise content taken from the research papers included in this dissertation. To improve the readability of the text, I omit the standard labeling of these citations.

integrated by way of technological development and advancement (Jeong et al., 2015). The Internet of Things (IoT) is one example that illustrates the difficulty of strict technology separation. Defined as smart devices connected to the internet and equipped with sensors, actuators, and intelligent computing logic (Bayer et al., 2021; Porter & Heppelmann, 2014), IoT has frequently been accorded an inherent intelligence, which has led to the notion of Artificial Intelligence (AI) embedded in IoT.

Technology convergence can lead to uncertainty about which combinations of technologies might cause the next disruptive innovation. Further ambivalence is created by the underlying accelerator of the megatrend that is digitalization. Compared to ancient technical innovations, such as the telephone, digital innovations nowadays take far less time to market, which explains the fast spread of digital technologies (Berger et al., 2018; Kose & Sakata, 2019). This rapidness makes it more difficult for humans to truly understand and keep up with each technological development. Especially as technologies become ever more complex, comprehension of their functioning gets more challenging. Even though their use does not require a profound understanding, at least not in most cases, a lack of transparency is likely to damage trust, which in turn may limit usage.

The challenge that comes with a lack of transparency is often discussed in the context of emerging technologies, particularly with AI. AI may be defined as a “system’s ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation” (Kaplan & Haenlein, 2019). The notion of the “black box” is frequently attributed to AI systems. It describes the opacity of its functions and outcomes for humans. With growing complexity, AI becomes less understandable for humans, which causes a feeling of uncertainty since it lies in human nature to be skeptical towards the unknown. One example is machine learning as subsymbolic AI. It underlines that, even though the calculations it requires are not complex, the sheer number of those calculations means that it is near enough impossible for its users and developers to truly comprehend or foresee its outcomes. The uncertainty about AI undermines trust – a crucial concept not only in human relationships but also in human-computer interaction. This becomes even more critical as AI enters tasks that bear risks for humans, such as autonomous driving, finance, or medicine (Barredo Arrieta et al., 2020). Especially in those critical contexts, there is a need to justify the system's outcomes (Barredo Arrieta et al., 2020). After all, since lack of trust is one of its significant limitations, and since this causes lower acceptance and usage (Miller, 2019), a popular solution approach is the concept of Explainable

AI (XAI). This approach aims to increase trust by providing the user with explanations to counteract the feeling that AI is a “black box” (Barredo Arrieta et al., 2020).

Besides missing transparency related to the so-called black box character of AI, critical perspectives on AI are often subsumed under the label of ethics. Both in practice and research, numerous ethical challenges are attributed to AI. One of the most famous examples related to ethics in AI is the moral machine, a “platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars” (MIT Media Lab, 2021). In a typical scenario, a person is driving a car and could save a child that is running on the street, but only if the car were to swerve to avoid a collision. In this case, the driver would die by crashing into a wall. Participants are asked what they think is the better decision. The moral machine gathers data about the moral values of its users, for instance, to draw cultural comparisons. This experiment refers to concerns about the potential differences between the moral values of users and the moral values embedded in the technology.

The often rather fuzzy definition of AI and its ethical issues relates in many contexts to the embedded algorithms (B. C. Stahl et al., 2021). Algorithms can influence which news are displayed in which sequence, they can define dynamic prices in online shopping, and they can preselect job applications (Diakopoulos, 2016; Martin, 2019; van den Broek et al., 2019). Decisions previously made by humans are increasingly supported or autonomously made by algorithms, which has raised widespread concerns about algorithmic decision-making (ADM). Among others, these concerns include discrimination as a result of biases in algorithms (Strobel, 2019), limited data privacy (Newell & Marabelli, 2015), and unclear responsibility when a decision has unintended consequences (Binns et al., 2018; Persson & Kavathatzopoulos, 2017). A self-driving car that provokes an accident could include reflections about the driver’s responsibility, the software engineer, the data scientist, the car itself, etc. The first recorded severe accident involving a self-driving car led to the death of Elaine Herzberg, who was hit by an Uber car in Arizona in 2018 (BBC, 2020). In this case, Uber was declared as “not criminal liable” for the accident, although the car appears to have failed to identify the victim as a pedestrian (BBC, 2019). Legalities aside, however, Uber paid a sum to the relatives of the deceased. The exact amount is not disclosed (Spiegel, 2018). The safety driver who was sitting in the car was charged with negligent homicide, because she was watching a TV show, instead of focusing on the street. The outcome of the trial is expected in August 2021 (BBC, 2019; Phoenix New Times, 2021). The intense discussions that have already taken place in the media with respect to the responsibility of Uber and the safety driver highlight the concerns which the public has about such unclear responsibilities. The same concerns may have a greater impact still on the extent to which the public will trust autonomous vehicles.

What all of these examples illustrate are current challenges associated with digitalization. The ever-greater reach of emerging technologies affects our lives in all areas. Indeed, an escape from technology is nigh on impossible these days. Even if one is not an active user, one's life is nonetheless likely to be touched by technological progress such as an encounter with an autonomous car or when one's job application is preselected by algorithms. Digitalization is not a new phenomenon, but what is new about the current wave is that the user has gained significant power to influence the direction of digitalization (Legner et al., 2017). Whereas digitalization used to focus on the professional environment, technologies have long since made their way into their users' private lives with users deciding if, when, and how they use technologies (Matt et al., 2019). In doing so, the individual user occupies various roles, such as that of the customer, the employee, or the individual itself (Matt et al., 2019). By means of such expectations, wishes, requests, or concerns, the user provides directions for innovation (Legner et al., 2017). This dissertation follows the idea of the central role of the individuum and focuses on the human perspective on emerging technologies, which includes, among others, opportunities for customers, individual user concerns, and solution approaches for software engineers.

Digitalization affects individuals in a range of positive ways – be it on a personal, an organizational, or a societal level, leading to new (digital) business models, products, and services (Legner et al., 2017). For example, IoT devices facilitate convenient shopping experiences, and algorithms generate highly personalized content due to the vast amount of collected data (Lee, 2015). Aside from such positives, however, certain perceived downsides cause users' reluctance towards emerging technologies and thus inhibit the adoption and proliferation of technology. Researchers investigate those negative side effects that impact individuals, organizations, or societies under the term "dark side of IS" (D'Arcy et al., 2014; Pirkkalainen & Salo, 2016; Tarafdar et al., 2015). Those negative phenomena can be either unexpected and unintended or deliberately provoked with malicious intent. Examples include cybercrime, technostress, technology addiction, or bias in algorithms (Majchrzak et al., 2016; Pirkkalainen & Salo, 2016; Turel & Serenko, 2012). A famous instance of a violation of data privacy is the Cambridge Analytica Scandal in which private Facebook data from tens of millions of users was illicitly acquired and abused to build voter profiles during the US presidential election campaign of 2016 (The New York Times, 2018). In media, there are many more examples of these adverse side effects of digitalization, underlining the interest and importance of this topic not only for research but also for practice. Examples include but are by no means limited to Google's algorithms that show higher-paid technical jobs to men, rather than to women, or those that classify black people as gorillas, while Amazon's machine learning systems have systematically downgraded CVs from, for instance, all-women schools (The Guardian, 2018a, 2018b; The Washington Post, 2015)

This dissertation covers the opportunities as well as the challenges that come with emerging technologies, but with a clear focus on the latter. Although the research field on the dark side of IS is not new, current IS research on opportunities by far outweighs research on the dark side (Gimpel & Schmied, 2019). Over the last years, this imbalance has been redressed due to a growing awareness of this deficit and the resulting calls for further research, e.g., from Pirkkalainen and Salo (2016), Tarafdar et al. (2015). However, there is still considerably less attention dedicated to the negative than to the positive aspects. Hence, with the spotlight on challenges, this dissertation counterbalances current research and enriches the examination of the dark side of IS.

1.2 Opportunities and challenges of emerging socio-technical systems

The open question remains how to ensure that the benefits of technology outweigh its adverse side effects on individuals, organizations, or societies. Since IS research is situated at the intersection of the technological artifact and the social context that develops or uses the technological artifact, IS research plays a pivotal role in finding answers to this question (Sarker et al., 2019). The common ground between the social and the technical is not strictly defined but rather a continuum between purely social and technical disciplines. Figure 1.2-1 illustrates this continuum of the social and technical disciplines.

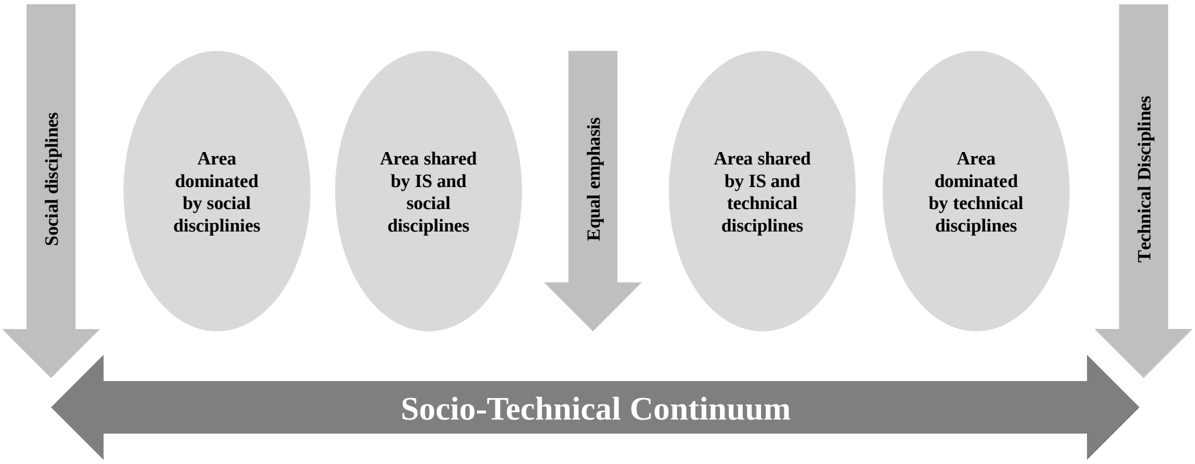

Social disciplines
Area dominated by social disciplinies
Area shared by IS and social disciplines
Equal emphasis
Area shared by IS and technical disciplines
Area dominated by technical disciplines
Technical Disciplines
Socio-Technical Continuum

Figure 1.2-1: The Socio-Technical Continuum (Sarker, 2019)

The left end of the continuum focuses on social disciplines and reduces technology to the context of examination, such as testing social theories in IT contexts (Sarker et al., 2019). The right end represents technical disciplines where social aspects recede into the background, for instance when priorities shift to the advancement of technical development. IS research should predominantly be positioned in between those two ends (Sarker et al., 2019). Within this focus falls the area dominated by social disciplines in which researchers are “treating technology as an outcome of social structure and processes” (Sarker et al., 2019, p. 702), with humans influencing technology. Next to the equilibrium, the social and the technical are considered decisive for a particular result, with the left-hand side seeing no interaction between the two aspects and the right-hand side focusing on the interplay between them. Finally, the area dominated by technical disciplines sees “technology as the major antecedent to social outcomes” (Sarker et al., 2019, p. 703) and includes research on how technology influences the social world.

During the last years, authors have stressed the need for IS research to shift the focus from either the technological or the social end towards a joint design of technologies and human systems, focus on interactions between technology, between technology and users (Lyytinen et al., 2020), equal consideration to the contextual factors and the environmental conditions of system use (Shin et al., 2014), and to take greater notice of ethical goals (Walsham, 2012). Most recently, Sarker et al. (2019) called for a return to the roots of the IS discipline, the socio-technical perspective that “considers the technical artifact as well as the individuals/collectives that develop and use the artifacts in social […] contexts” (Sarker et al., 2019, p. 696). Following Sarker et al. (2019), this dissertation aims at a “healthy distribution” (p. 708) of papers along the socio-technical continuum.

Along the continuum, researchers can indulge in opportunities and challenges of socio-technical systems. Following the model of Vial (2019), the use of digital technologies (e.g., social, mobile, IoT, platforms, ecosystems) facilitates changes in value creation paths (e.g., digital channels, value proposition), which can have positive as well as negative results. The positive aspects include greater organizational efficiency, increased organizational performance, or improvements in healthcare (Agarwal et al., 2010; Vial, 2019). Meanwhile, the adverse outcomes comprise, for instance, extensively researched problems surrounding security and privacy (Vial, 2019). As Sarker et al. (2019) have pointed out, IS research has typically focused on instrumental goals, rather than on humanistic goals. Compared to IT, the whole information system, not only the technology, is at the core of examination in IS research (Lee, 2015). It is with this in mind that humanistic goals, such as equality and well-being, ought to be included in IS research alongside instrumental outcomes, such as efficiency and productivity. In view of the strong interdependence, equally considering technical and social components along the socio-technical continuum is expected to meet both instrumental and humanistic goals better than a focused study of one side alone (Bostrom et al., 2009; Sarker et al., 2019). The current spotlight on desirable outcomes of technology has eclipsed research on dark side phenomena. A growing share of publications specifically focuses on dark side phenomena (e.g., Gimpel & Schmied, 2019; Kim et al., 2011; Pirkkalainen & Salo, 2016), but as mentioned above, the dark side of IS remains a minor research area compared to the expansive work done on the opportunities of IS. This dissertation takes these insights as a point of departure to indulge in opportunities, but in particular, challenges of IS to counterbalance the current surplus in research about opportunities of digitalization.

Furthermore, this dissertation specifically focuses on emerging technologies. Digital technologies comprise emerging technologies such as AI, IoT, and blockchain as well as established technologies, such as social media platforms (Berger et al., 2018). It is one defining characteristic of digitalization that the emergence and adoption of new technologies are fast, in private or professional life (Berger et al., 2018). Hence, emerging technologies are highly dynamic in the sense that technologies often develop quickly from the development phase to the market phase. Gartner provides an annual Hype Cycle of Emerging Technologies, naming current emerging technologies and classifying them according to their maturity level, for instance, the innovation trigger, the peak of inflated expectations, or a plateau of productivity (Gartner, 2021).

Along with technologies such as the health passport, private 5G, and social distancing technologies, the Hype Cycle of 2020 includes a remarkable number of AI technologies, such as AI-assisted design, composite AI, responsible AI, embedded AI, and explainable AI. Moreover, explainable AI and embedded AI are at the peak of the Hype Cycle, whereas all other AI technologies are still in the first phase, also known as the innovation trigger (Gartner, 2020a). In sum, the multiplicity of AI-related technologies in the Hype Cycle illustrates the current importance of this technology, which in turn illustrates its specific relevance to further examination in this dissertation.

Next to AI, IoT played a central role in the Hype Cycles of the past years, which has led to the notion of a “hyper-connected world” (Shin et al., 2014). Kevin Ashton introduced the term IoT in 1999. His vision was a world in which every physical object is connected to the Internet via ubiquitous sensors (Shin et al., 2014). The Hype Cycle of Emerging Technologies of 2020 no longer uses the term IoT but inhibits the IoT-related technology digital twin (Gartner, 2020a). Next to the general Hype Cycle of Emerging Technologies, Gartner publishes industry-specific and technology-specific Hype Cycles. The Hype Cycle for IoT shows that its first technologies peaked over the last few years and are now entering the disillusion phase. This includes technologies like IoT platform and IoT security (Gartner, 2020c). Other IoT technologies are still in the first phase of the cycle and are expected to peak in the coming years. These include technologies such as things as Customers and IoT-Enabled Products as a Service (Gartner, 2020c). The Hype Cycle for Supply Chain Strategy expects IoT is two to five years away from impact in the industry, as IoT has already been implemented in many companies, but an efficient and productive usage of its opportunities has yet to be defined (Gartner, 2020b; Modern Materials Handling, 2020). Many current discussions, both in practice and research, center on the usefulness of IoT for humans. The fact that it is possible does not necessarily mean that it is useful. This distinction, however, is rarely made during IoT development, and the focus often lies on technological possibilities (Shin et al., 2014). To counterbalance this trend, this dissertation covers the opportunities as well as challenges of IoT.

In sum, the Hype Cycles show the underlying dynamic in emerging technologies that lead to fast changes, not only in their allocation along the Hype Cycle but also in the fragmentation of technology into multiple phenomena related to the respective technology. This dissertation focuses on AI and IoT as two emerging technologies that have both been established in the Hype Cycle and are widely discussed in practice and research.

1.3 Aim and outline of this dissertation

At present, there is a surplus of literature on the opportunities afforded by technology, and this is supplemented by a rather small share of research on its challenges. As Sarker (2019) observed, IS research should benefit society. Accordingly, the motivation of IS research should be human-driven, rather than purely technology-driven. It should aim at a balanced distribution of IS research along the socio-technical continuum, whereas current IS research focuses on the ends of this continuum. This dissertation takes these insights as a point of departure. It aims to advance knowledge for IS, with regard to opportunities, and in particular with a focus on challenges of emerging technologies along the socio-technical continuum. More specifically, this dissertation focuses on the emerging technologies AI and IoT. The framework of Sarker (2019) is used to structure the research papers of this cumulative dissertation along the socio-technical continuum. This dissertation does not cover the entire breadth of the socio-technical continuum but aims instead at a “healthy distribution” of research papers along this axis of cohesion that provides guidance for the communication of knowledge in IS research (Sarker et al., 2019).

Table 1.3-1 provides an overview of the structure of this dissertation, including brief summaries of all research articles included. For each article the titles, objectives, methods, and co-authors are provided. Chapter 2 refers to the examination of IoT, chapter 3 to that of AI.

Chapter 2: Behind the scenes of IoT				
Section	**Title of the research paper**	**Objective**	**Method**	**Co-Authors**
2.1	IoT-commerce: Opportunities for customers through an affordance lens	Identifying and structuring opportunities of IoT devices for retail commerce customers.	Structured literature review for the identification of affordances; validation regarding completeness and parsimony with a sample of 337 IoT devices	Gimpel, Henner Rau, Daniel
2.2	IoT ethics – Status quo and directions for further research	Identifying and structuring IoT ethics and providing guidance for further research.	Structured literature review for the identification of ethical issues of IoT	-
Chapter 3: Behind the scenes of AI				
Section	**Title of the research paper**	**Objective**	**Method**	**Co-Authors**
3.1	Fear of algorithms: A synopsis of concerns about automated decision-making	Identifying, structuring, and communicating individual concerns about ADM to improve ADM-related offers and services that consider the perspectives of individuals.	Structured literature review with qualitative content analysis of semi-structured interviews	Schmied, Fabian Waldmann, Daniela
3.2	Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence	Discussing whether we can grasp the descriptive and normative dimensions of AI and especially sub-symbolic machine-learning-based systems with the help of ‘agency’ attribution.	Argumentative deductive analysis	Fritz, Alexis Brandt, Wiebke Gimpel, Henner
3.3	The role of domain expertise in trusting and following explainable AI decision support systems	Examining how the domain-specific expertise of human users influences their trust in explainable AI decision support systems and their behavior regarding going along with the system’s suggestions.	Online experiment with a self-developed AI system	Gimpel, Henner Markgraf, Moritz
3.4	Towards a systematic inclusion of ethical impacts in design and development of software: A Framework for Ethical Software Development	Developing an approach for ethical software development that provides guidance on how software development teams should incorporate the software product’s potential ethical impacts during the design and development process.	Design Science Research based on a literature review and semi-structured interviews	Fähnle, Annika Gimpel, Henner

Table 1.3-1: Overview of the research articles of the dissertation

Chapter 2 of this dissertation focuses on the emerging technology IoT. Section 2.1 addresses the opportunities of IoT situated at the left side of the socio-technical continuum, followed by a synopsis of IoTs' ethical issues which cover the entire bandwidth of the continuum in section 2.2. In the following, more details about the two research articles related to IoT are provided.

From a customer's perspective, the opportunities afforded by IoT are specifically present in the context of commerce. After the breakthrough and establishment of electronic and mobile commerce, IoT is expected to radically transform purchasing cultures (J. Shim et al., 2019). Well-discussed examples in literature and media include the smart self-ordering fridge and the voice assistants with the help of whom one can order products (Evans, 2017; Rothensee, 2008). However, the literature on IoT mainly has focused on the technical features of this technology, leaving aside the customer perspective. To fill this research gap, section 2.1 answers the following research question: *"Which opportunities do Internet of Things devices provide to retail commerce customers?"* Due to the massive size of the retail sector and the disruptive potential of IoT, answering this research question is valuable in that it links research on IoT with knowledge of commerce through the lens of the customer perspective, offering guidance for further research and to practitioners who wish to improve the design of customer experience in retail commerce. To answer the research question, IoT-commerce is analyzed through an affordance lens embedded in Activity Theory.

After the examination of opportunities afforded by IoT, section 2.2 focuses on research on IoT ethics. Given the enormous potential of IoT, it is expected to continue its spread into multiple areas of our life (Shim et al., 2020). While first ideas have begun to materialize, for the most part, we are still merely seeing early signs of IoT's potential. Actual impacts will only become apparent in the future (Avital et al., 2019). To realize its full potential, it is crucial to address potential ethical issues and concerns that may be associated with IoT, since resolving them is of key significance to its acceptance and spread. Literature on IT and ethics is primarily tailored to AI, yet due to the specific characteristics of IoT, general transferability of those issues to IoT is doubted, which is to say that the literature does not dedicate sufficient attention to IoT ethics (Cascone et al., 2017). To fill this gap, *this paper aims to identify and structure ethical issues of IoT discussed in literature and connects the issues with IoTs' features and illustrate them with exemplary application contexts.* It discusses the current state of research, identified through a structured literature review, and proposes directions for further research.

Chapter 3 of this dissertation focuses on AI as an emerging technology and includes four research papers situated along the socio-technical continuum. First, the concept of AI is analyzed through the lens of the concerns that individuals have about automated decision-making, targeting the roots of AI, namely the underlying algorithms embedded in AI systems (section 3.1). In the middle of the continuum, one specific concern about AI, unclear responsibility, is analyzed through an ethical perspective, including the concept of moral agency (section 3.2). In line with the necessity to address individuals' concerns towards AI systems, users have to trust an AI system to ensure its use. Trust in AI systems, often referred to as black-box systems, is not self-evident. With this in mind, section 3.3 dives into the concept of XAI and examines the role of domain expertise in trusting AI decision support systems and acting in accordance with their advice. Finally, section 3.4 proposes an ethical software development framework. driven by the idea that to address individuals' concerns, ethical issues have to be embedded from the beginning of a software development process, chapter 3.4 proposes an ethical software development framework. The following paragraphs provide further details about each of the research articles related to AI.

The paper situated furthest towards the social end of the continuum examines concerns about ADM (section 3.1). Decisions previously made by humans are increasingly supported by algorithms, ranging from simple queries to AI (Martin, 2019; Wachter et al., 2017). Application areas of ADM are diverse, ranging from recommender systems for online shopping to calculating recidivism rates in court (Angwin et al., 2016). As algorithms are getting more complex, their outcomes become less understandable and traceable for users and software engineers (Westin et al., 2016). This lack of transparency brings concerns about the use of automated decision-making. In the literature, these concerns are discussed in specific use cases of ADM, but to date, there is no comprehensive overview of concerns held by individuals. However, knowing and understanding potential concerns about ADM is crucial for its adoption, which is why this paper aims to answer the following research question: *"Which concerns do individuals have about the use of automated decision-making?"* The answer is developed out of a structured literature review and semi-structured interviews about concerns in multiple ADM use cases.

Situated in the middle of the continuum, the concept of moral agency in the context of AI is examined in section 3.2. The paper argues that, although philosophers and sociologists are increasingly attributing agency to AI, the concept of agency should solely be attributed to human agents. Three ethical models of human-computer interaction from Floridi (Floridi, 2016; Floridi & Sanders, 2001, 2004), Johnson and Verdicchio (Johnson & Verdicchio, 2018), and Verbeek (Verbeek, 2006, 2011, 2014, 2017) are analyzed based on explanations of symbolic and sub-symbolic AI, the network around machine learning, and the agency concept discussed in teleology-naturalism and actor-network theory.

To the right of the equilibrium, the role of domain expertise in trusting and following explainable AI decision support systems is analyzed in section 3.3. User trust is of critical importance to further use of AI in various application areas (Biran & Cotton, 2017). However, since human beings can only comprehend AI to a limited extent, users tend to be reluctant to embrace the unknown (Biran & McKeown, 2017). XAI tries to overcome this obstacle by offering the user explanations of AI's outcomes (Biran & Cotton, 2017). So far, there is no blueprint of what constitutes a good explanation. This paper examines the role played by a user's domain-specific expertise when it comes to setting up explanations, aiming to answer the following research question: *"How does the domain-specific expertise of human users influence their trust in explainable AI decision support systems (XAI DSS) and their behavior in regard to going along with the systems' suggestions?"* Furthermore, the paper examines the influence on actual behavior, which is to say whether or not the user truly follows the advice of the AI. Hypotheses are tested with data from an online experiment as well as an associated survey for each participant.

In section 3.4, a framework for ethical software development is proposed, the purpose of which is to include ethical impacts in the design and development of software. Due to a lack of transparency on the one hand and ethical significance on the other, software development ought to account for potential ethical impacts from the beginning of the process (Allen et al., 2006; Spiekermann & Winkler, 2020). Therefore, *the paper aims to "develop an approach for ethical software development, named ethical software development process model (ESDP), that provides guidance on how software development teams should incorporate the software product's potential ethical impacts during the design and development process."* This framework is built on design science research, including a literature review and semi-structured expert interviews.

Figure 1.3-1 embeds the research articles included in this dissertation along the socio-technical continuum. The columns of the matrix show the socio-technical continuum, while the rows show the two emerging technologies that are examined in this dissertation, AI and IoT. Each dimension is characterized by opportunities and challenges, a selection of which is examined in this dissertation.

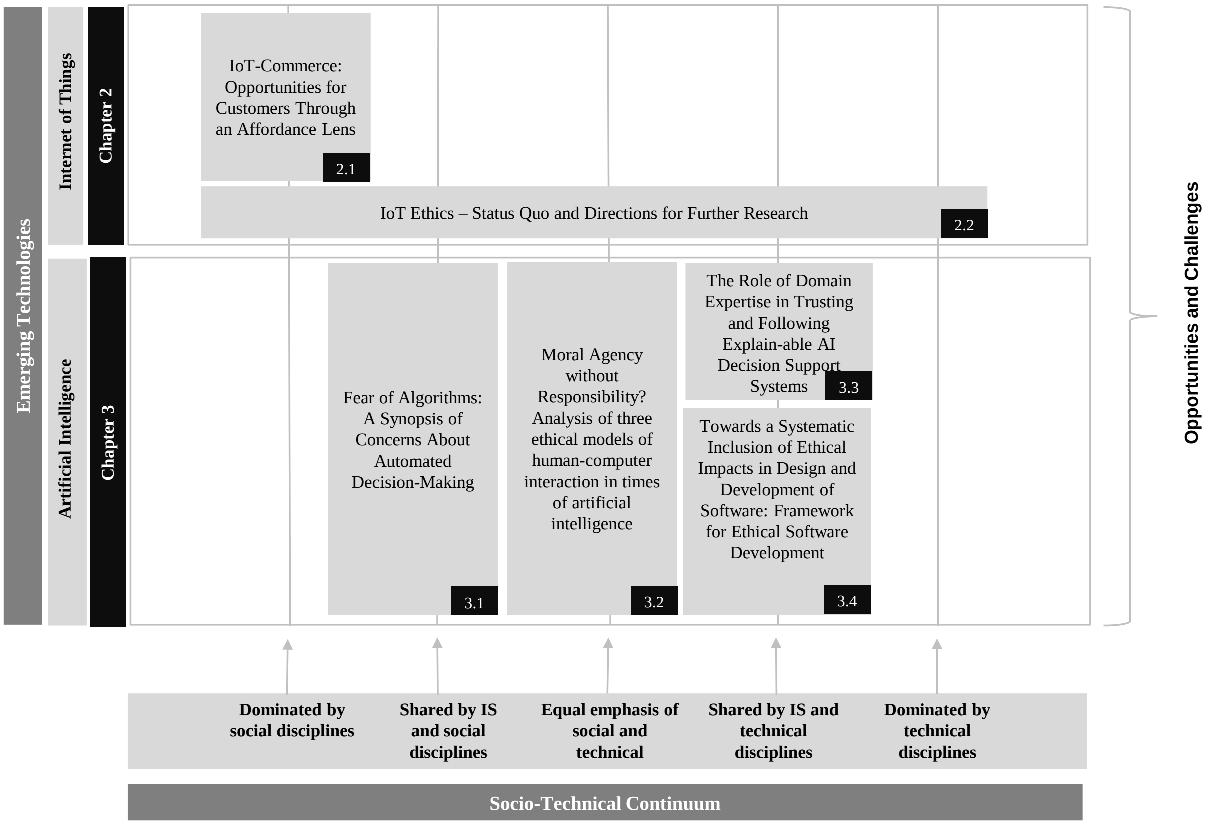

Emerging Technologies
Internet of Things
Chapter 2
IoT-Commerce: Opportunities for Customers Through an Affordance Lens
2.1
IoT Ethics – Status Quo and Directions for Further Research
2.2
Artificial Intelligence
Chapter 3
Fear of Algorithms: A Synopsis of Concerns About Automated Decision-Making
3.1
Moral Agency without Responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence
3.2
The Role of Domain Expertise in Trusting and Following Explain-able AI Decision Support Systems
3.3
Towards a Systematic Inclusion of Ethical Impacts in Design and Development of Software: Framework for Ethical Software Development
3.4
Opportunities and Challenges
Dominated by social disciplines
Shared by IS and social disciplines
Equal emphasis of social and technical
Shared by IS and technical disciplines
Dominated by technical disciplines
Socio-Technical Continuum

Figure 1.3-1: Structure of this dissertation along the socio-technical continuum

These research articles are prefaced by an introduction (Chapter 1) that includes the motivation of this dissertation, provides its theoretical base, and describes its outline. The research articles are then followed by a discussion and conclusion (Chapter 4), which includes a summary of the results and implications, points ahead at opportunities for further research, and presents an overall conclusion of this dissertation.

References

Agarwal, R., Guodong, G., DesRoches, C., & Jha, A. K. (2010). The digital transformation of healthcare: Current status and the road ahead. *Information Systems Research*, *21*(4), 796–809.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, *21*(4), 12–17.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 29.12.2019.

Avital, M., Dennis, A. R., Rossi, M., Sørensen, C., & French, A. (2019). The transformative effect of the internet of things on business and society. *Communications of the Association for Information Systems*, *44*, 129–140.

Barblan, B. (2018). *Technology should serve humanity. Not the other way round.* https://www.nothing.ch/en/research/technology-should-serve-humanity-not-other-way-round. Accessed 16.07.2021.

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

Bayer, S., Gimpel, H., & Rau, D. (2021). Iot-commerce - opportunities for customers through an affordance lens. *Electronic Markets*, *31*(1), 27–50.

BBC. (2019). *Uber 'not criminally liable' for self-driving death*. https://www.bbc.com/news/technology-47468391. Accessed 20.06.2021.

BBC. (2020). *Uber's self-driving operator charged over fatal crash*. https://www.bbc.com/news/technology-54175359. Accessed 20.06.2021.

Bednar, P. M., & Welch, C. (2020). Socio-technical perspectives on smart working: Creating meaningful and sustainable systems. *Information Systems Frontiers*, *22*(2), 281–298. https://doi.org/10.1007/s10796-019-09921-1

Berger, S., Denner, M., & Röglinger, M. (2018). The nature of digital technologies – development of a multi-layer taxonomy. *Proceedings of the 26th European Conference on Information Systems (ECIS), Portsmouth, UK.*

Binns, R., van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). It's reducing a human being to a percentage. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* Montreal QC, Canada.

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *Xai workshop at the 26th international joint conference on artificial intelligence,* Melbourne, Australia.

Biran, O., & McKeown, K. (2017). Human-centric justification of machine learning predictions. In *26th international joint conference on artificial intelligence,* Melbourne, Australia.

Bostrom, R. P., Gupta, S., & Thomas, D. (2009). A meta-theory for understanding information systems within sociotechnical systems. *Journal of Management Information Systems*, *26*(1), 17–48.

Cascone, Y., Ferrara, M., Giovannini, L., & Serale, G. (2017). Ethical issues of monitoring sensor networks for energy efficiency in smart buildings: A case study. *Energy Procedia, 134*, 337–345.

D'Arcy, J., Gupta, A., Tarafdar, M., & Turel, O [O.] (2014). Reflecting on the “dark side” of information technology use. *Communications of the Association for Information Systems, 35*(5), 109–118.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56–62. https://doi.org/10.1145/2844110

Evans, M. (2017). *5 ways the internet of things will influence commerce*. https://www.forbes.com/sites/michelleevans1/2017/01/24/5-ways-the-internet-of-things-will-influence-commerce. Accessed 30.01.2019.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences, 374*(2083).

Floridi, L., & Sanders, J. W. (2001). Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*, *3*, 55–66.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*, 349–379.

Gartner. (2020a). *5 trends drive the gartner hype cycle for emerging technologies 2020*. https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020/. Accessed 15.06.2021.

Gartner. (2020b). *Gartner 2020 hype cycle for supply chain strategy shows internet of things is two to five years away from transformational impact*. https://www.gartner.com/en/newsroom/press-releases/2020-09-09-gartner-2020-hype-cycle-for-supply-chain-strategy-shows-internet-of-things-is-two-to-five-years-away-from-transformational-impact. Accessed 14.06.2021.

Gartner. (2020c). *Hype cycle for the internet of things, 2020*. https://www.gartner.com/en/documents/3987602/hype-cycle-for-the-internet-of-things-2020. Accessed 14.06.2021.

Gartner. (2021). *Gartner hype cycle: Interpreting technology hype*. https://www.gartner.com/en/research/methodologies/gartner-hype-cycle. Accessed 14.06.2021.

Gimpel, H., & Röglinger, M. (2015). *Digital transformation: Changes and chances – insights based on an empirical study*. http://fim-rc.de/Paperbibliothek/Veroeffentlicht/542/wi-542.pdf. Accessed 28.05.2021.

Gimpel, H., & Schmied, F. (2019). Risks and side effects of digitalization: A multi-level taxonomy of the adverse effects of using digital technologies and media. *Proceedings of the 27th European Conference on Information Systems (ECIS2019), Stockholm-Uppsala, Sweden*.

The Guardian. (2018a). *Amazon ditched ai recruiting tool that favored men for technical jobs*. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine. Accessed 20.06.2021.

The Guardian. (2018b). *Google's solution to accidental algorithmic racism: Ban gorillas*. https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people. Accessed 20.06.2021.

Jeong, S., Kim, J.-C., & Choi, J. Y. (2015). Technology convergence: What developmental stage are we in? *Scientometrics, 104*(3), 841–871.

Johnson, D. G., & Verdicchio, M. (2018). Ai, agency and responsibility: The vw fraud case and beyond. *AI & SOCIETY*, *34*(3), 639–647.

Kaplan, A., & Haenlein, M. (2019). Siri, siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25.

Kim, W., Jeong, O. R., Kim, C., & So, J. (2011). The dark side of the internet: Attacks, costs and responses. *Information Systems*, *36*(3), 675–705.

Kose, T., & Sakata, I. (2019). Identifying technology convergence in the field of robotics research. *Technological Forecasting and Social Change*, *146*, 751–766. https://doi.org/10.1016/j.techfore.2018.09.005

Lee, J. K. (2015). Research framework for ais grand vision of the bright ict initiative. *Management Information Systems Quarterly*, *39*(2), iii-xii.

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.

Lyytinen, K., Nickerson, J. V., & King, J. L. (2020). Metahuman systems = humans + machines that learn. *Journal of Information Technology*.

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. *Management Information Systems Quarterly*, *40*(2), 267-277.

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

Matt, C., Trenz, M., Cheung, C. M. K., & Turel, O [Ofir] (2019). The digitization of the individual: Conceptual foundations and opportunities for research. *Electronic Markets*, *29*(3), 315–322.

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

MIT Media Lab. (2021). *Moral machine*. https://www.moralmachine.net/. Accessed 03.06.2021.

Modern Materials Handling. (2020). *Gartner's hype cycle: Iot in "trough" but transformational stage on the way*.

https://www.mmh.com/article/gartners_hype_cycle_iot_in_trough_but_transformational_stage_on_the_way. Accessed 27.05.2021.

The New York Times. (2018). *Cambridge analytica and facebook: The scandal and the fallout so far*. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html. Accessed 17.06.2021.

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems*, *24*(1), 3–14. https://doi.org/10.1016/j.jsis.2015.02.001

Persson, A., & Kavathatzopoulos, I. (2017). How to make decisions with algorithms: ethical decision-making using algorithms within predictive analytics. *ACM Computers & Society*, *47*(4).

Phoenix New Times. (2021). *Trial delayed for backup driver in fatal crash of uber autonomous vehicle*. https://www.phoenixnewtimes.com/news/uber-crash-arizona-vasquez-herzberg-trial-negligent-homicide-charge-11553424. Accessed 20.06.2021.

Pirkkalainen, H., & Salo, M. (2016). Two decades of the dark side in the information systems basket : Suggesting five areas for future research. *Proceedings of the 24th European Conference on Information Systems (ECIS), Tel Aviv, Israel*.

Porter, M. E., & Heppelmann, J. E. (2014). How smart, connected products are transforming companies. *Harvard Business Review*. https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition

Rachinger, M., Rauter, R., Müller, C., Vorraber, W., & Schirgi, E. (2019). Digitalization and its influence on business model innovation. *Journal of Manufacturing Technology Management*, *30*(8), 1143–1160.

Rothensee, M. (2008). User acceptance of the intelligent fridge: empirical results from a simulation. In *Proceedings of the 1st international conference on the internet of things (iot 2008)* (pp. 123–139).

Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the is discipline: Its historical legacy and its continued relevance. *Management Information Systems Quarterly*, *43*(3), 695–719.

Shim, J., Avital, M., Dennis, A. R., Rossi, M., Sørensen, C., & French, A. (2019). The transformative effect of the internet of things on business and society. *Communications of the Association for Information Systems, 44*(1), 129-140.

Shim, J. P., Sharda, R., French, A. M., Syler, R. A., & Patten, K. P. (2020). The internet of things: Multi-faceted research perspectives. *Communications of the Association for Information Systems, 46*, 511–536.

Shin, D.-H., Yoon, H., Lee, J., Moon, Y., Kim, N., & Hoyeon, C. (2014). A socio-technical framework for internet-of-things design. *20th Biennial Conference of the International Telecommunications Society (ITS), Rio De Janeiro, Brazil*.

Spiegel. (2018). *Tod durch algorithmus*. https://www.spiegel.de/panorama/tod-durch-algorithmus-a-8b012d8d-0002-0001-0000-000161218857. Accessed 23.06.2021.

Spiekermann, S., & Winkler, T. (2020). Value-based engineering for ethics by design. *Preprint ArXiv:2004.13676*. https://arxiv.org/abs/2004.13676

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Laulhé Shaelou, S., Patel, A., Ryan, M., & Wright, D. (2021). Artificial intelligence for human flourishing – beyond principles for machine learning. *Journal of Business Research*, *124,* 374–388.

Strobel, M. (2019). Aspects of transparency in machine learning: doctoral consortium. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019),* Montreal.

Tarafdar, M., Bolman, E., & Ragu-Nathan, T. S. (2015). Technostress: Negative effect on performance and possible mitigations. *Information Systems Journal*, *25*(2), 103–132.

Tarafdar, M., Gupta, A., & Turel, O [O.] (2015). Special issue on 'dark side of information technology use': An introduction and a framework for research. *Information Systems Journal*, *25*(3), 161–170.

Turel, O [Ofir], & Serenko, A. (2012). The benefits and dangers of enjoyment with social networking websites. *European Journal of Information Systems, 21*(5).

van den Broek, E., Sergeeva, A., & Huysman, M. (2019). Hiring algorithms: An ethnography of fairness in practice. In *Proceedings of the 40th International Conference on Information Systems (ICIS),* Munich, Germany.

Verbeek, P.-P. (2006). Materializing morality. Design ethics and technological mediation. *Science, Technology, & Human Values, 31,* 361–380.

Verbeek, P.-P. (2011). Moralizing technology. Understanding and designing the morality of things. *Chicago: Univ. Of Chicago Press*.

Verbeek, P.-P. (2014). Some misunderstandings about the moral significance of technology. *The Moral Status of Technical Artefacts, Edited by Peter Kroes and Peter-Paul Verbeek. Dordrecht: Springer*, 75–88.

Verbeek, P.-P. (2017). Designing the morality of things: The ethics of behaviour-guiding technology. *Designing in Ethics, Edited by Jeroen Van Den Hoven, Seumas Miller and Thomas Pogge. New York: Cambridge Univ. Press*, 78–94.

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *Journal of Strategic Information Systems Review*, *28*(2), 118–144.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Walsham, G. (2012). Are we making a better world with icts? Reflections on a future agenda for the is field. *Journal of Information Technology*, *27*(2), 87–93.

The Washington Post. (2015). *Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you.* https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/. Accessed 20.06.2021.

Westin, C., Borst, C., & Hilburn, B. (2016). Automation transparency and personalized decision support: Air traffic controller interaction with a resolution advisory system. *IFACPapersOnLine*, *49*(19), 201–206.

2 Behind the scenes of the Internet of Things

2.1 IoT-commerce: Opportunities for customers through an affordance lens

Abstract:

Retail commerce is influenced by digital technologies at large scale. After electronic commerce and its evolution into mobile commerce, we now see that the Internet of Things (IoT), one of the most disruptive developments in recent times, is about to radically transform retail commerce from need recognition to post-purchase engagement and service. Extant literature mainly investigates technical features of IoT, missing out on a customer-centric perspective. Theoretically founded in Activity and Affordance Theories, this paper conceptualizes IoT-commerce, identifies opportunities for customers, and links them to the customer buying process. Based on an extensive literature review, twelve affordances are derived and evaluated with a sample of real-world IoT devices. All affordances offered by electronic and mobile commerce are still valid for IoT-commerce but extended by three affordances unique to IoT-commerce: context-aware services, natural interactions, and automated customer processes. Affecting all steps of the customer buying process, IoT-commerce is worth to be understood by researchers, customers, and companies.

Keywords: activity theory, affordances, e-commerce, m-commerce, IoT

Authors: Sarah Bayer, Henner Gimpel, Daniel Rau

Status: This article is published in Electronic Markets – The International Journal on Networked Business, Volume 31, Issue 1, pp. 27-50 (2020).

2.1.1 Introduction

As one of the most disruptive developments in recent times, the Internet of Things (IoT) has the power to radically transform retail commerce (Shim et al., 2019). The application of IoT in retail commerce might even be “the most profound shift ushered in by the IoT era” (Evans, 2018, p. 1). IoT refers to a multitude of smart devices that are connected to the Internet and equipped with sensors, actuators, and intelligent computing logic (Porter & Heppelmann, 2014). “In time, the idea of a smartphone as a commerce device could be old news as commerce moves beyond simply portable consumer devices to include durable goods, such as refrigerators, washing machines or automobiles.” (Evans, 2017, p. 1). Taking the example of a smart fridge as an IoT device, first publications were discussing it already more than ten years ago (Coughlan et al., 2012; Gaur et al., 2015; Rothensee, 2008). “The smart fridge has often been considered a prototypical example of applications of the Internet of Things” (Rothensee, 2008, p. 123). Exemplary functionality comprises the tracking of expiry dates, recipe recommendations, and an automated re-ordering of groceries almost used up. Similar use cases of IoT within retail commerce are described for washing machines automatically re-ordering detergent (Deloitte, 2016) or pet food dispensers (Amazon, 2018b). Only recently, however, such ideas began to materialize. For instance, Samsung is now offering a smart fridge with a built-in touchscreen that allows adding items to the shopping list and then directly order products online (Groenfeldt, 2016). Walmart filed a patent for a technology that would allow automatic re-purchase of groceries and other products without any further intervention by the customer (Nassauer, 2017), an IoT idea that is already implemented in a similar way for detergent in the washing machines of Whirlpool and GE Appliances (Evans, 2017). With tens of millions of sold devices, Amazon’s (2018a) voice-controlled Echo is a popular example of an adopted IoT device already widely used in retail commerce (Reid, 2018). This smart speaker connects to other compatible smart home devices and its ecosystem integration with the Amazon marketplace delivers seamless online shopping experiences into the homes of customers that allow the purchase of products with simple voice commands. Hence, being only theoretically discussed for many years, we now see first materialized examples of IoT in the context of retail commerce. We call this phenomenon IoT-commerce.

Ever since, commerce refers to the activity of buying and selling and, therefore, to the exchange of tangible and intangible goods at large scale (Oxford Dictionary, 2018). For centuries, brick and mortar stores represented a common way of retail commerce. Driven by new technologies, new forms of commerce evolved. Increased penetration of the Internet led to the opportunity to sell and buy products online using webshops and electronic data transmission (Grandon & Pearson, 2004). Revenue of this global electronic commerce (e-commerce) is anticipated to triple from 1.34 trillion USD in 2014 to 4.13 trillion USD in 2020 (Statista, 2018b). With double-digit growth rates of around 21% annually, e-commerce contributes substantially to the growth of global retail sales. After the turn of the millennium, the proliferation of mobile Internet-enabled smartphones facilitated spatially independent access to online shopping (Clarke, 2008). In a 2017 survey, every third online shopper stated to purchase online via a mobile device at least once per month (Statista, 2018a). Along these lines, mobile commerce (m-commerce) created an unparalleled opportunity as it expanded the traditional limitations of e-commerce (Clarke, 2008). Nowadays, as examples such as smart fridges and voice assistants show, we see those limitations expanding once again, driven by IoT-commerce.

Extant literature at the intersection of IoT and commerce is scarce. IoT literature mainly takes a technology-centered perspective, such as describing functionalities and applications of IoT devices (Borgia, 2014), groups of IoT devices (Püschel et al., 2016), the related technology stack (Porter & Heppelmann, 2014), interaction patterns between IoT devices, customers, and businesses (Kees et al., 2015; Oberländer et al., 2018), and the interplay of different stakeholders in smart service systems (Porter & Heppelmann, 2014). Existing commerce literature mainly focuses on e-commerce and m-commerce.

There exists few IoT literature including customer perspectives, such as a discussion about how companies can enhance customer value with IoT devices via energy savings, property protection, proactivity, or personalized experience (Koverman, 2016; Lee & Lee, 2015). Nevertheless, we are not aware of a study taking the device itself as a starting point for holistically describing what IoT devices afford to customers with regard to commerce. Yet, due to the relevance of retail commerce, the substantial changes that came along with e-commerce and m-commerce, and the disruptive potential of the IoT for retail commerce, we posit that such a perspective is valuable and a prerequisite for further examinations of IoT-commerce. Therefore, we raise the research question:

Which opportunities do Internet of Things devices provide to retail commerce customers?

Answering this research question is valuable from a theoretical and a practical perspective. In terms of theory, it links the rather disparate bodies of knowledge on commerce and IoT in a customer-centric manner. This is a basis for future research exploring the growing field of IoT-commerce. From a practical perspective, answering the research question is relevant for customers using IoT devices for shopping and companies that might want to (re-)design their customer experience in light of a new channel for customer interaction.

To answer our research question, we analyze IoT-commerce through an affordance lens embedded in Activity Theory. We use Activity Theory as a meta-theoretical lens, and we leverage the general and information systems-specific affordance literature as a more specific theoretical foundation. Activity theory provides the socio-economic framework of rules (e.g., legislation), community (e.g., other customers), and roles (e.g., socially discriminating factors) as “minimal context for individual actions” (Beaudry & Carillo, 2006, p. 429), that means the purchasing of products and services online through the means of IoT devices. Affordance Theory helps us describe individual opportunities for customers that emerge with the use of IoT devices in the context of retail commerce. Our work grounds on both academic literature in the fields of e-commerce, m-commerce, and IoT as well as on real-life examples of IoT devices that allow the purchasing of products and services. E-commerce and m-commerce literature is included in our assessment in order to evaluate whether their affordances are still valid or even strengthened in IoT-commerce. IoT literature is included in order to identify additional affordances offered by IoT devices that are not yet present since e-commerce and m-commerce. As a validation for parsimony and completeness, we derive potential manifestations of the affordances within the different steps of the customer buying process and consider real-life examples of IoT devices to assess the extent of actual manifestations and check for completeness and parsimony. Our analysis revealed twelve affordances of IoT in the context of retail commerce that manifest in all steps of the customer buying process.

Section 2 provides the theoretical background on the evolution of e-commerce, m-commerce, and IoT-commerce and our theoretical foundation in Activity and Affordance Theories. Section 3 outlines the methodological approach. Section 4 presents the affordances of IoT-commerce from a theoretical perspective. A validation with real-life objects is described in section 5. Section 6 presents the discussion, followed by the conclusion in section 7.

2.1.2 Theoretical background

In the following, we introduce the customer buying process on which our paper is based. Subsequently, we explain the evolution of commerce in three waves since the emergence of the World Wide Web in the 1990s. Thereby, we provide an overview of the development and the existing literature streams of e-commerce, m-commerce, and IoT. Afterward, we outline Activity Theory and Affordance Theory as theoretical foundations of our work.

2.1.2.1 Customer Buying Process

The customer buying process is a series of activities where a customer interacts in several stages with a seller or manufacturer. Models of Howard and Sheth (1969), Nicosia and Mayer (1976) and Engel et al. (1995) are often-cited models of buyer behavior including a large number of constructs, as, for instance, word of mouth, perceptual bias, and intention to purchase. Those models are taken from numerous publications as the basis for examining the influence of individual variables, such as perceived risk (Cunningham et al., 2005), or experience of the decision-maker (Frambach et al., 2007).

Our paper aims to detect opportunities of IoT within the buying process rather than to describe the underlying psychological process of the decision-maker. Therefore, we searched for a well-structured buying process focusing on the core of the process rather than on influencing variables. We selected the customer buying process of Lemon and Verhoef (2016) as it offers clear and distinguishable steps for our analysis of IoT's influence in each of those steps. The steps of this process, published in one of the leading business journals, are depicted in Figure 2.1-1 and explained afterward.

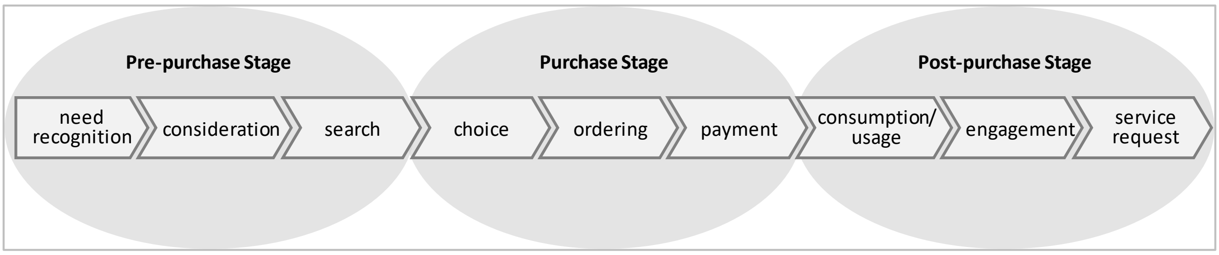


Figure 2.1-1: Customer buying process adopted from Lemon and Verhoef (2016)

The customer buying process of Lemon and Verhoef (2016) differentiates three major stages: pre-purchase, purchase, and post-purchase. In the pre-purchase stage, the customer interacts with other parties, as, for instance, the brand, without beginning the actual purchase. Afterward, she recognizes a need for something (*need recognition*), she considers buying something to satisfy this need (*consideration*), and she gathers information via searching product alternatives (*search*). In the purchase stage itself, the customer chooses one alternative (*choice*), orders the product or service (*ordering*), and pays for it (*payment*). The post-purchase stage summarizes the potential interaction of the customer with the brand or the environment related to the product or service after the actual purchase. The product or service is consumed or used (*consumption/usage*). Furthermore, the customer can evolve into some sort of post-purchase engagement (*engagement*) and send service requests concerning the product or service (*service* *request*) (Lemon & Verhoef, 2016). Not every buying process follows this pattern chronologically; steps can also be swapped or omitted (e.g., buying something without searching for alternatives, or engaging with the manufacturer before consuming a product).

Numerous publications take this buying process as a starting point, often to examine one or more steps in detail. Especially the steps *search* and *choice* are considered as critical steps in the process and are therefore often subject to examination in e-commerce and m-commerce literature. For instance, Brynjolfsson and Smith (2000) disproved the assumption that lower search costs on the Internet lead to a preference of lower-priced stores instead of higher-priced stores for the same product. Trust in the online retailer (reputation, word of mouth, advertising, or prominent links from other trusted websites) or brand loyalty were found to make customers choose higher-priced products (Brynjolfsson & Smith, 2000; Kocas, 2002). Lynch and Ariely (2000) argue that e-commerce reduces the search costs for products and product-related information, leading to the suggestion for retailers to offer differentiated goods. An investigation into search costs in context of IoT-commerce could be valuable, as customers may even outsource the *search* partially or completely to the IoT device, for instance, by ordering the cheapest product via voice, the product of a specific retailer, or by ordering unspecific and letting the IoT device decide based on customer's preferences. All those scenarios lead to lower search costs in IoT-commerce than in e-commerce and m-commerce that might make it even more crucial for companies to either build trust, offering highly differentiated products, or focus on price competition. Similarly, product comparison agents assist customers in decision-making leading to decreased search costs in e-commerce and unlimited availability of alternatives (Wan et al., 2007). Among different types of product comparison agents, such as evaluation agents, differentiation agents, or preference agents, IoT-commerce might especially enhance preference agents as the gathering of the underlying customer preference data significantly increases with IoT devices (e.g., shopping history, interactions with the device, music preferences, daily routines, etc.).

Based on the above-mentioned customer buying process, we now study the phenomenon of IoT-commerce as an evolution of e-commerce.

2.1.2.2 The Evolution of Electronic Commerce

Above all, e- commerce is "a technologically driven phenomenon" (Laudon & Traver, 2018, p. 38), subject to constant change due to technological advancements (Ngai & Gunasekaran, 2007; Strader & Shar, 1997). The emergence of new technological devices that were adopted by a critical number of private users and that inhibit the possibility to support each step of the customer buying process or only parts of it influenced the way retail commerce is conducted. We see three major waves of commerce initiated via the widespread use of new technologies: e-commerce, enabled by desktop devices (e.g., Personal Computer) connected to the Internet; m-commerce, enabled by mobile devices (e.g., smartphones and tablets) connected to the Internet; and IoT-commerce, enabled by IoT devices (e.g., voice assistants) connected to the Internet.

Before the debut of mail-order catalogs and teleshopping, brick and mortar stores were the linchpin of buying for all kinds of products and services (Miles, 1990). With the appearance of the World Wide Web in the early 1990s, the base for e-commerce was set (Turban et al., 2015). E-commerce "suggests that consumers access a website through a computer terminal" (Maity & Dass, 2014, p. 35). The literature discusses several aspects that distinguish e-commerce from traditional commerce. Electronic marketplaces, for example, electronic payment or electronic marketing via social networks, allow the "participating buyers and sellers to exchange information about prices and product offerings" (Alt & Klein, 2011; Strader & Shar, 1997, p. 187; Turban et al., 2015). When O'Reilly Media first used the term 'Web 2.0' in 2004, they described the evolution of the World Wide Web toward social media, facilitating information sharing between customers, for instance, in social networks or by writing reviews (Butler & Peppard, 1998; Turban et al., 2015). Along these lines, increased information density allows customers to instantly acquire detailed product information. Additionally, e-commerce enables customization of products and personalization of services (Butler & Peppard, 1998; Turban et al., 2015).

Academic literature in the field of e-commerce is rich, comprising topics like technical aspects (e.g., Guttman et al., 1999; Lee & Lee, 1993; Xiao & Benbasat, 2007), behavioral issues such as consumer behavior and technology acceptance (e.g., Gefen et al., 2003; Klopping & McKinney, 2004; Liang et al., 2011), and business models (e.g., Aldridge, 1998; Kraemer et al., 2000; Timmers, 1998). Several literature reviews synthesize extant e-commerce research streams from different perspectives. For instance, Ngai and Wat (2002) screened 275 articles published between 1993 and 1999 and clustered them into four categories: 'application areas,' 'technological issues,' 'support and implementation,' and 'others'. Chua et al. (2005) took a stakeholder perspective and identified 'customers' and the 'internal organization' as stakeholders with the most attention in e-commerce research, whereas 'suppliers', 'indirect stakeholders', 'investors', and 'regulators' receive less interest. In their review of electronic markets research, Alt and Klein (2011) identified three perspectives: 'economic environment,' 'governance mode,' and 'business model'.

With the advent of more and more mobile devices in the 2000s, m-commerce took its course (Wirtz, 2018). M-commerce describes the possibility "to purchase goods anywhere through a wireless Internet-enabled device" (Clarke, 2008, p. 133; Maity & Dass, 2014). Mobile devices in the context of m-commerce are portable devices with wireless Internet access that, by nature, are designed to be moved with its users, such as smartphones and tablets (Junglas & Watson, 2003; Turban et al., 2015). Furthermore, mobile devices enable location-based services and advertisement that is individually adapted to the local context. For instance, the Uber platform allows to call a taxi to the current GPS location of the customer and estimates its time of arrival (Turban et al., 2015). Although some functions, such as user accounts for desktop PCs and synchronization of user accounts on different devices, offer strong personalization possibilities in e-commerce, personalized advertising in m-commerce is typically easier as mobile devices are used by only one person (Turban et al., 2015).

Mobile devices experienced their upswing after the emergence of desktop PCs for e-commerce. However, in some regions m-commerce unfolded first, followed by later adaption of e-commerce. Furthermore, we observe that the beginning of e-commerce is product-oriented, meaning that products of brick-and-mortar stores were bought via the Internet, whereas the ongoing change from product-orientation towards service-orientation initiated a change in e- and m-commerce towards services. Some characteristics of e-commerce expand in m-commerce, and some key attributes of m-commerce, such as location-based services, lead to specialized business models, especially in the service domain (Turban et al., 2015). In our paper, we examine products, that means "something that is made to be sold, usually something that is produced by an industrial process" (Cambridge Dictionary, 2019a), services, that means "business activity that involves doing things for customers rather than producing goods" (Cambridge Dictionary, 2019b), as well as hybrid offers of products and services.

Similarly to e-commerce, research streams of m-commerce relate to technical aspects (e.g., Y. E. Lee & Benbasat, 2004), behavioral research (e.g., Schierz et al., 2010), and business models (e.g., Tsalgatidou & Pitoura, 2001). However, the total number of publications is significantly smaller compared to e-commerce. M-commerce research focuses on the additional features provided by mobile devices compared to desktop devices, for instance, location-based services (e.g., Rao & Minakakis, 2003). There are only a few literature reviews of m-commerce. For instance, Ngai and Gunasekaran (2007) synthesize m-commerce research streams by the categories of ‘m-commerce theory and research’, ‘wireless network infrastructure’, ‘mobile middleware’, ‘wireless user infrastructure’, and ‘m-commerce applications and cases’. Groß (2015) clustered m-commerce literature into the three categories of ‘online distribution channel’, ‘advanced technology for in-store shopping’, and ‘technology perspective’.

The overview of existing research about e-commerce and m-commerce shows that authors frequently take a technical point of view (e.g., technical aspects of e-commerce/m-commerce, technological issues, wireless network infrastructure, support and implementation, mobile middleware), which is not the focus of our paper. Within our research, we concentrate on publications that describe commerce from a customer’s point of view or in such a general way, that implications for customers can be derived. Furthermore, we assume that the research streams of technical aspects, behavioral research, and business models that flourished through the waves of e-commerce and m-commerce are likely to be continued for IoT-commerce. Our research on IoT-commerce is located at the intersection of all three research streams as we use an affordance lens to derive opportunities from technical features of IoT devices that impact customer behavior and potentially enable innovative business models.

With the emergence of IoT devices, also called smart connected devices, another option to purchase online arises next to desktop devices (e-commerce) and mobile devices (m-commerce). IoT describes the phenomenon that physical objects are integrated into the networked society, leading to a fusion of the physical and digital world (Huber et al., 2017; Rosemann, 2014; Wortmann & Flüchter, 2015). Two central aspects turn devices into IoT devices. First, the Internet connection, enabling the device to send and receive data. Secondly, sensors and/or actuators enable those objects to be “tracked, coordinated, or controlled across a data network or the Internet” (McKinsey, 2013, p. 52). The IoT device equipped with sensing and acting capabilities captures and aggregates data, and potentially takes action (Borgia, 2014). Therefore, the IoT device possesses certain intelligence to act and make decisions independent of human agency (Gaskin et al., 2014; Porter & Heppelmann, 2014). In short, IoT connects information technology and physical objects, leading to new products and services (McKinsey, 2013; Uckelmann et al., 2011). Having reviewed different definitions of IoT (Huber et al., 2017; McKinsey, 2013; Uckelmann et al., 2011; Wortmann & Flüchter, 2015), we find the definition of Kees et al. (2015) most suitable for this paper as it gives a good understanding of IoT from the viewpoint of the user, which is in our context the customer. In line with Kees et al. (2015) we define:

IoT devices are a multitude of physical objects, equipped with sensors, actuators, and/or computing power connected to the Internet via communication technology, and enabling interaction with and/or among those objects.

Personal computers, laptops, tablets, and mobile phones are traditional physical devices with sensors, computing power, and typically an Internet connection. These devices are IoT devices. However, commerce solely relying on the aforementioned devices commonly used in e-commerce and m-commerce does not qualify as IoT-commerce. Rather, IoT-commerce denotes retail commerce using non-traditional smart connected physical devices such as voice assistants, smart washing machines, and smart thermostats. With the help of voice assistants, one can purchase new products within seconds via voice command; with a smart washing machine, one can automatically reorder detergent right before it is used up; with a smart thermostat, temperature and therefore energy consumption can be optimized automatically without assistance of the owner leading to an adjusted amount of energy purchase, especially in cases where heating and cooling is performed with electricity. This implies that IoT-commerce is not a radical replacement of e-commerce and/or m-commerce, but can be seen as an “evolution rather than a revolution” (Evans, 2017, p. 1). Hence, we use the following definition of IoT-commerce:

IoT-commerce relates to the purchasing of products and services online via the use of IoT devices whose technical features afford new opportunities to retail customers.

Information systems research discusses features, interactions, and recommendations of IoT devices (e.g., Fleisch et al., 2009; Oberländer et al., 2018; Song et al., 2017). In marketing, interactions between the IoT device and the customer are discussed related to customer and object experiences (e.g., Hoffman & Novak, 2018; Kozinets, 2019). Under the terms “ubiquitous commerce”, computer scientists examine the ubiquity and pervasiveness of IoT-commerce from a technical and opportunity-centered perspective, describing potential applications of IoT in commerce (e.g., Bhajantri et al., 2015; Chunxia et al., 2010; Fox et al., 2006; Sanchez-Pi & Molina, 2009). Examples of IoT devices used in retail commerce are already widely discussed in practice (e.g., Farhad, 2018; Heatman, 2018), but scientific research embedded in existing theories is still scarce in this field.

2.1.2.3 Activity Theory

Our work of IoT devices in retail commerce builds on Activity Theory, which is located at the intersection of behavioral and social sciences, as it “provides a high-level contextual perspective of human behavior" (Beaudry & Carillo, 2006, p. 429). Activity Theory formalizes the interaction of a subject with the world. The conceptualization of Activity Theory, especially regarding activity systems of individuals, traces back to Vygotsky (1980) and Leont'ev (1978) and is also known as the mediated-action perspective (Kaptelinin & Nardi, 2012). In particular, it describes how a Person (P) interacts with an Object (O) via the use of Tools (T) (Benbunan-Fich, 2019). The Person is typically an acting human being. The Object is affected by the action of the Person (Benbunan-Fich, 2019). Kaptelinin (2005) differentiates between physical objects and intangible constructs such as commonly accepted facts or socially and culturally defined properties that can also represent the Object within Activity Theory. The Object, therefore, possesses an ambiguous nature (physical vs. intangible). Ultimately, the Object is the motive for the actions of a goal-directed Person (Kaptelinin, 2005). The Tool mediates the human activity of the Person, contributes to accomplishing the intended goal, and triggers an effect on the Object. The nature of Tools can be both physical such as technology (Karanasios & Allen, 2014) or psychological such as language, symbols, and mental models (Allen et al., 2013; Karanasios & Allen, 2014). Another typology of tools by Hasan and Kazlauskas (2014) differentiates between primary (physical), secondary (language, mental models), and tertiary tools (communities, context, or environment). Typically, the relationship between Person, Tool, and Object is depicted as 'P ⬄ T ⬄ O'. Later, Engeström (1987) extended this triad to incorporate the socially embedded concepts of rules, community, and roles (Beaudry & Carillo, 2006).

As a well-articulated concept for descriptive purposes, Activity Theory is a theory for analysis and explanation of the world (Benbunan-Fich, 2019; Gregor, 2006) with activities as the unit of analysis (Engeström, 1987). In its original conceptualization by Vygotsky (1980) and Leont'ev (1978), Activity Theory did not particularly comprise digital technologies. Against this backdrop, researchers introduced Activity Theory into the domain of Human-Computer Interaction to better understand how technology mediates human activities (Kaptelinin & Nardi, 2012; Nardi, 1996). Until now, two major Information Systems (IS) research streams evolved around Activity Theory, one to better understand IS intervention and the other "more connected with the fields of design and development, and the technical side of IS" (Benbunan-Fich, 2019, p. 3). As a cross-disciplinary framework, Activity Theory contributes to a human-oriented understanding of the collaboration and interaction between humans (i.e., Person) and IS (i.e., Tool). It allows investigating different types of human practices on both an individual and social level (Nardi, 1996).

IS research has started to utilize Activity Theory in the domain of e-commerce. For instance, Chaudhury et al. (2001) built their work on Activity Theory to understand customer experiences in the Internet and to support successful web development. Johnston and Gregor (2000) rely on core elements of Activity Theory to conceptualize industry-level activity that aims at explaining certain aspects of supply chain e-commerce technologies. Beaudry and Carillo (2006) organize their review of B2C literature along with the Activity Theory framework. In this paper, we build on Activity Theory as a foundation to explore the affordances of IoT devices in the customer buying process. In the words of Activity Theory, our goal-directed Person is the customer with her goal to satisfy her need via online purchase. By the use of an IoT device as a primary physical Tool, this customer interacts with a seller, manufacturer, or service provider that can be seen as Object. Her interaction is embedded in the socio-economic framework of rules (e.g., legislation), community (e.g., peer customers), and roles (e.g., social demography).

2.1.2.4 Affordance Theory

Repeatedly, Activity Theory is combined with the concept of affordances as they are the relational property of interaction within the Person-Tool- Object triad (Benbunan-Fich, 2019). Whereas instrumental affordances relate to the handling (‘P ⇔ T’) and effect (‘T ⇔ O’) of the Tool, supplemental affordances relate to auxiliary activities such as maintenance of the Tool (Benbunan-Fich, 2019). The concept of affordances was first brought up in ecological psychology. It originates from Gibson (1979) who used the verb ‘afford’ to describe what the environment offers to an animal. He refers to a subject (animal) that is provided with affordances from an object (object within the environment). Gibson (1979) associates properties of objects with affordances that guide the actions of the subject. He, therefore, emphasizes the complementarity of a subject and its environment (Benbunan-Fich, 2019). Later, Norman (1988) introduced affordances into design theory and the domain of Human-Computer Interaction. In the beginning, his work centered mainly on his design-oriented belief that objects and tools should be designed for their intended use – in a way that the user can anticipate the object’s affordance. He then abstracted from the physical nature of objects and applied the concept of affordances on intangible artifacts and software user interfaces (Benbunan-Fich, 2019). Therefore, he rather took a design perspective on affordances. In further work, Norman (1999) differentiated between perceived and real affordances. Whereas real affordances refer to the actual properties of an object or artifact, the perceived affordances are those that are noticeable for subjects such as human beings by providing cues for proper operation and usage. Depending on the individual, perceived affordances may vary among a heterogeneous group of users.

Based on the classification by Norman (1999), there were attempts to further extend the classification of affordances. Hartson (2003), for instance, suggested a differentiation between cognitive (i.e., perceived affordances), physical (e.g., real affordances), sensory (i.e., properties to feel, see hear, etc.), and functional (i.e., support in a task relating to a higher purpose) affordances. Vyas et al. (2017) conceptualize affordances at a much broader scope by incorporating social and cultural aspects. However, most authors interpret the relationship between subject and object as the core of the Affordance Theory (Bærentsen & Trettvik, 2002; Benbunan-Fich, 2019; Gaver, 1991, 1992; Gibson, 1979; Kaptelinin & Nardi, 2012; McGrenere & Ho, 2000; Norman, 1999). Affordances materialize in the interaction between the subject (e.g., Person/human) and object (e.g., IoT device). Following this logic, affordances are possibilities for goal-directed actions of goal-oriented actors with regards to an object (Markus & Silver, 2008). The affordance perspective can, therefore, provide a useful lens to analyze (emerging) technologies in a user-centered manner (Gaver, 1991; Leonardi, 2011).

In the following, we rely on the general concept of affordances but do not further differentiate between different types such as cognitive, physical, sensory, and functional affordances of IoT-commerce. We use the term ‘IoT-commerce affordances’ for affordances of IoT devices directed to retail commerce customers and apply the same analogy on ‘e-commerce affordances’ and ‘m-commerce affordances’. Furthermore, our affordances might not yet all be perceived by customers. Hence, we focus on real affordances for now, though future research on the differences to perceived affordances might be very useful. Perceived affordances might then be actualized by the customer within the process of purchasing products and/or services online.

2.1.3 Methods

To answer our research question which opportunities IoT devices provide to retail commerce customers, we pursued a two-step approach in which theory development is followed by validation. For theory development, we identified the affordances of e-commerce, m-commerce, and IoT-commerce based on academic literature. For validation, we conducted a twofold analysis to ensure parsimony and completeness. Below, we provide details on each methodological step.

In the first step of theory development, we reviewed extant literature to collect real affordances of e-commerce, m-commerce, and IoT-commerce. As literature on the established research domains of e-commerce and m-commerce is rich, we focused on articles synthesizing existing research. By contrast, the IoT phenomenon is not yet well researched. Therefore, we did not further restrict our search to literature review articles about IoT but conducted our search in the whole IoT domain. We used the following combined search term for titles and abstracts: `{{"e-commerce" OR "electronic commerce" OR "m-commerce" OR "mobile commerce" OR "online shopping" OR "electronic shopping" OR "mobile shopping" OR "e-business" OR "electronic business"} AND {review OR affordance}} OR {iot OR "internet of things"}`. As advised by Webster and Watson (2002), we performed our literature search in leading IS journals, namely the AIS Senior Scholars' Basket of Eight (2018). Furthermore, following Webster and Watson (2002), we expanded our search beyond core IS journals. We included the journal 'Electronic Markets' due to its inherent connection to electronic commerce, and other peer-reviewed journals specifically addressing the electronic commerce domain[6]. Furthermore, we integrated a marketing perspective due to its close connection to commerce. We searched in leading marketing journals[7] for `"internet of things" OR "IoT"`. To include discussions about IoT in computer science and electrical engineering[8], we additionally searched for the corresponding terms: `{"e-commerce" OR "electronic commerce" OR "m-commerce" OR "mobile commerce" OR "online shopping" OR "electronic shopping" OR "mobile shopping" OR "e-business" OR "electronic business"} AND {"ambient intelligence" OR "pervasive computing" OR "ubiquitous computing"}`.

The search resulted in 180 articles on which two authors independently performed a title and abstract screening. An article was considered relevant if it mainly dealt with e-commerce, m-commerce, or IoT, provided an overview of the evolution of at least one of those fields, or presented one specific affordance in detail. An article was marked for detailed examination if at least one researcher classified it as relevant. With this research strategy within IS journals, commerce-related journals, marketing journals and the domain of computer science, we "accumulate a relatively complete census of relevant literature" (Webster & Watson, 2002, p. xvi).

As a second step, two researchers independently examined the full text of the remaining 49 relevant journal articles in detail and highlighted affordances of e-commerce, m-commerce, and

[6] Considered journals: Electronic Commerce Research, Electronic Commerce Research & Applications, International Journal of Electronic Commerce, Journal of Electronic Commerce in Organizations, Journal of Electronic Commerce Research, Journal of Organizational Computing & Electronic Commerce

[7] Considered journals: Journal of Marketing, Journal of Marketing Research, Journal of Consumer Research

[8] Database used: ieeexplore.ieee.org

IoT-commerce. We identified phrases (e.g., ‘24/7 availability’), sentences, passages, and the whole topic of an article (e.g., ‘user-generated content in the form of online product reviews’) as affordance if they satisfied the following criterion: The aspect is peculiar to e-commerce, m-commerce, or IoT-commerce, therefore helps identify the respective phenomenon, and offers direct or indirect possibility for action to the customer. For each affordance, we documented its presence within e-commerce, m-commerce, and/or IoT-commerce. Relatively few publications on IoT in the context of commerce revealed that IoT-commerce affordances cannot be compiled solely on commerce-related literature. Hence, during the paper screening, we also highlighted (technical) features and aspects of the very nature of IoT devices that lead to affordances for customers.

In intense discussions, we consolidated all aspects highlighted during the paper screening, finally leading to twelve affordances as presented in the next section. Within this consolidation, same and similar highlighting was merged into one affordance (e.g., ‘24/7 availability’ and ‘temporal independence’), the granularity level of all affordances was harmonized (i.e., not too specific and not too generic affordances), an explanation comprising all relevant aspects identified in academic literature was compiled, and the number of affordances was decreased to achieve conciseness.

For validation with real-life IoT devices regarding completeness and parsimony of our theory, we chose a twofold approach. We drew a sample of 337 IoT devices that were obtained from three studies that provide extensive literature reviews of IoT devices in scientific and grey literature: Oberländer et al. (2018), Püschel et al. (2016), and Brandt et al. (2017). For our research, we considered only those IoT devices that either enable the purchase of products and services by itself (e.g., Amazon Echo) or strongly influence the type, quality, quantity, or ordering time of goods purchased (e.g., Nest thermostat), resulting in 35 relevant IoT devices. Therefore, other devices such as smart locks (e.g., Lockitron), smart mattresses (e.g., Luna), or smart home monitoring systems (e.g., Sentri) were not considered. Similar IoT devices (e.g., Amazon Echo Dot and Amazon Echo Plus or Nest thermostat and smart irrigation controllers) were grouped, resulting in five major groups of these 35 IoT devices relevant for IoT-commerce as presented by Table 2.1-1. The overview contains devices for the purchase of both products and services. Further, those purchases are made either explicitly (i.e., purchase immediately initiated by the device) or implicitly (e.g., optimization of energy consumption which transitively influences the amount of energy purchased).

#	Group of IoT devices	Short explanation	Purchase target	Purchase behavior	Examples (selected)	Examples in our sample
1	Voice assistants	Voice assistants are applications built on natural language processing that allows the customer to interact with the device in her native language. The voice assistant can carry out requests in interaction with other connected devices and services. For instance, the voice assistant controls music, sets alarms, makes audio calls, or buys products and services online. Typical voice assistants are implemented in smart speakers but also watches and even car assistance systems.	Product and services	Explicit	Amazon Echo, iMCO Watch	6
2	Smart resource management	To optimize energy consumption, devices for smart resource management use different kinds of data such as location, customer preferences, sensor data like temperature, or weather forecasts to automatically adjust the consumption of water, energy, and other resources to environmental conditions. For most devices, the user can monitor and control the consumption via a built-in interface, via mobile applications, or via a connected voice assistant.	Product	Implicit	Nest thermostat, Skydrop	18
3	Replenishment services	Replenishment services allow to (re-)order products based on customer's preferences – explicitly stated from the customer or implicitly detected through her purchase history. Customers can re-order products by pressing a button, by scanning the barcode of any available product, or without any intervention when the replenishment service decides autonomously on products to (re-)order. Replenishment services may also be implemented in devices such as smart fridges, smart washing machines, etc.	Product	Explicit	Amazon Dash Wand, Samsung Family Hub	5
4	Rental services	Bike and car sharing systems allow customers to easily rent vehicles via a mobile app. Locations of vehicles and their availability are accessible through a mobile app or stationary terminals as all vehicles are connected to the Internet. Cars or bikes are activated by wireless key solutions. For billing, the time and usage locations of the vehicles are tracked and then charged via the rental platform the customer has registered for.	Service	Implicit	DriveNow, Smoove	4
5	Maintenance services	Modern cars or other vehicles and devices are equipped with maintenance services. Automatic notifications are sent to manufacturers or dealers when the vehicle or device needs maintenance. Based on this trigger, the customer is contacted to agree on the time and scope of the maintenance service. Further processing such as the payment may be handled through the same platform.	Service	Explicit	GM OnStar Dealer Maintenance Notification	2

Table 2.1-1: Categories of IoT devices

To check for parsimony, we evaluated whether all twelve identified affordances do already manifest in reality. We derived manifestations of affordances by subsequently applying each affordance to each step of the customer buying process. If the examined real-life examples of IoT devices confirmed that an affordance provides a possibility for action to the customer or removes her need to act in one of those buying process steps, we documented the respective manifestation and provided an exemplary description of this opportunity directed to the customer. Our overview of affordances presented in the next section only comprises IoT-commerce affordances that already manifest in several steps of the customer buying process as supported by existing real-life examples of IoT devices.

To check for completeness, we chose the reverse approach and examined real-life examples of IoT devices in detail. For each device, we analyzed its influence on each step of the customer buying process. If an IoT device has an influence on one step of the buying process and provides an immediate opportunity to the customer in this step, then we checked whether this opportunity is already covered by the identified twelve affordances. As all opportunities provided by real-life IoT devices were already covered by the IoT-commerce affordances we identified, we did not have to add further affordances.

The following section 4 presents twelve affordances of IoT-commerce as a result of the theory development, followed by validation with real-life objects in section 5.

2.1.4 Affordances of IoT-commerce

The main result of our paper is the identification of affordances for IoT-commerce as shown in Table 2.1-2. For each affordance, we provide a definition based on the aspects raised in literature. Furthermore, we state with a bullet ('●') in the second, third, and fourth column in which wave of commerce an affordance takes effect. Interestingly, no affordance disappeared from one wave to the following ones. All affordances that emerged with e-commerce can still be found in the following two waves and all affordances that emerged with m-commerce are also offered by IoT-commerce. Some affordances that are presented to only occur within the second or third wave can also have an effect on the previous waves. For instance, the extensive automation of customer processes due to smart algorithms primarily emerged with the new opportunities of data collection by IoT devices (e.g., by their sensors), but is now also reactively influencing e-commerce and m-commerce. However, in this section, we focus on IoT-commerce affordances and their respective origins and do not mark retrospective effects on e-commerce and m-commerce. All affordances based on the 49 relevant articles of our literature review are presented with justificatory references.

Affordance	e-com.	m-com.	IoT-com.	Justificatory references
[1] Electronic transactions Business transactions related to the buying and selling of products and services are conducted partially or fully via the Internet. Steps of the buying process previously handled offline are carried out online. This change in transaction patterns leads to convenient and novel shopping experiences.	●	●	●	Akter & Wamba, 2016; Hollocks, 2001; Levina & Vilnai-Yavetz, 2015; Nan et al., 2017; Pousttchi et al., 2015; Romano & Fjermestad, 2002; Sharma & Gutiérrez, 2010; Song et al., 2017; Vaithianathan, 2010; Xu & Gutierrez, 2006
[2] Temporal independence Shopping is independent of temporal restrictions, as electronic transactions can be carried out 24/7, without the need to consider opening hours.	●	●	●	Akter & Wamba, 2016; Fleisch et al., 2009; Lehrer et al., 2018; Levina & Vilnai-Yavetz, 2015; Nan et al., 2017; Pousttchi et al., 2015; Samaras, 2002; Sanchez-Pi & Molina, 2009; Sharma & Gutiérrez, 2010; Song et al., 2017; Vaithianathan, 2010; Xu & Gutierrez, 2006
[3] Online platforms Online marketplaces with virtual storefronts offer a broad range of products and services. These platforms emerge as market and distribution intermediaries that aggregate the supply of one or many manufacturers or retailers. Some platforms even work without any professional seller as they connect peer customers with each other.	●	●	●	Akter & Wamba, 2016; Gengatharen & Standing, 2005; S. M. Lee et al., 2011; Liu et al., 2016; Song et al., 2017
[4] Information transparency Large amounts of information are published online. Devices provide (ubiquitous) access to information for customers but also for third parties, to details such as on products and manufacturers, marketplace reviews, and shopping experiences.	●	●	●	Adolphs & Winkelmann, 2010; Kurkovsky, 2005; Liu et al., 2016; Song et al., 2017; Vaithianathan, 2010; Xiao & Benbasat, 2007
[5] Social interactions User-generated content can easily be created and shared online, for instance via social networks or product reviews in online shops. This enables new ways for customers to communicate with manufacturers, retailers, and peers. The amount of information shared between customers increases as customers value user-generated content as a trustful source of information. This so-called electronic word-of-mouth enhances the online shopping experience.	●	●	●	Al-Obeidat et al., 2018; Baek et al., 2012; Baethge et al., 2016; Cui et al., 2018; Fu et al., 2018; Ho et al., 2017; Huang & Benyoucef, 2013; Konjengbam et al., 2018; Leong et al., 2016; Levina & Vilnai-Yavetz, 2015; Ma et al., 2017; Manvi et al., 2011; Mengxiang et al., 2017; Nan et al., 2017; Peng et al., 2016; Ramaswamy & Ozcan, 2018; Safi & Yu, 2017; Saumya et al., 2018; Song et al., 2017; Yan Wan et al., 2018
[6] Personalized services Based on data about the customer and the customer's context, the steps of the buying process are tailored to a specific customer or a group of customers. This leads to individualized services as well as customized products tailored to a customer's personal profile.	●	●	●	Adolphs & Winkelmann, 2010; Akter & Wamba, 2016; Baethge et al., 2016; Bhajantri et al., 2015; Fox et al., 2006; Huang & Benyoucef, 2013; Jing et al., 2018; Kurkovsky, 2005; Lehrer et al., 2018; S. Li & Karahanna, 2015; Liu et al., 2016; Nan et al., 2017; Prasad, 2003; Shang et al., 2012; Sharma & Gutiérrez, 2010; Song et al., 2017; Xiao & Benbasat, 2007; Xu & Gutierrez, 2006
[7] Proactive services Based on data about the customer and her context, automated trigger-based action is independently carried out by a system. For instance, the customer automatically receives a recommendation without having actively asked for it.	●	●	●	Akter & Wamba, 2016; Barbosa, 2015; Lehrer et al., 2018; S. Li & Karahanna, 2015; Shang et al., 2012; Song et al., 2017; Xu & Gutierrez, 2006
[8] Spatial independence Shopping is independent of spatial restrictions. With their ubiquitous nature, portable mobile devices allow the customer to purchase products and services via wireless networks from any location.		●	●	Chunxia et al., 2010; Fox et al., 2006; Nan et al., 2017; Samaras, 2002; Sanchez-Pi & Molina, 2009; Shang et al., 2012; Sharma & Gutiérrez, 2010; Song et al., 2017; Xu & Gutierrez, 2006
[9] Location-based services Based on tracking previous and current locations and foreseeing future locations of the customer, specific location-based actions can be triggered.		●	●	Fleisch et al., 2009; Kurkovsky, 2005; Pousttchi et al., 2015; Sharma & Gutiérrez, 2010; Xu & Gutierrez, 2006
[10] Context-aware services User context and situational context of the customer and her products are utilized to provide related information and services. The ubiquitous acquisition of context data with sensors and the transmission of data with actuators is followed by intelligent reasoning and a suitable (real-time) reaction to it. Thereby, existing value propositions for customers can be improved and novel services can be created.			●	Barbosa, 2015; Chunxia et al., 2010; Fleisch et al., 2009; Fox et al., 2006; Hoffman & Novak, 2018; Kurkovsky, 2005; Xuemei Li et al., 2008; Manvi et al., 2011; Sanchez-Pi & Molina, 2009; Shang et al., 2012; Song et al., 2017
[11] Natural interactions Smart devices come along with new user interfaces. The customer can interact naturally (e.g., via voice, haptics, gesture) with pervasive IoT devices during the customer buying process. This leads to a broader view of value creation and new customer experiences through interactions in the interactive systems environment.			●	Fleisch et al., 2009; Fox et al., 2006; Kozinets, 2019; Kurkovsky, 2005; Ramaswamy & Ozcan, 2018; Sanchez-Pi & Molina, 2009
[12] Automated customer processes The whole buying process or parts of it are conducted automatically by algorithms without the customer having to interact with someone or something. Decisions that were previously made by the customer are now automated, increasing convenience and saving time for the customer. This is possible due to extensive data collection via connected IoT devices and machine-to-machine communication.			●	Fleisch et al., 2009; Fox et al., 2006; Kurkovsky, 2005; Xuemei Li et al., 2008; Oberländer et al., 2018; Song et al., 2017

Table 2.1-2: Affordances of e-commerce, m-commerce, and IoT-commerce

The twelve affordances are split into seven that primarily arose with e-commerce, two affordances that primarily emerged with m-commerce, and three affordances that experienced their upswing with IoT-commerce. IoT-commerce itself is characterized by all of the twelve affordances as those of previous waves still remain valid. However, the sheer number of affordances might suggest that e-commerce is the most important wave of commerce. Though the Internet accessible through desktop computers resulted in a substantial change in retail commerce, IoT-commerce also holds disruptive potential. A rapidly growing number of connected devices and first manifestations of IoT affordances in retail commerce demonstrate that IoT-commerce is about to radically transform the way online purchases are made. Hence, an investigation into the affordances of IoT-commerce is truly valuable. The three affordances that distinguish IoT-commerce from the previous waves of commerce result from technical features of IoT devices such as described by the editorial of Fleisch et al. (2009). Our paper goes beyond the described technical features, applies IoT functionality to the context of retail commerce, and incorporates recent developments and more specific concepts of the IoT phenomenon since the publication of Fleisch et al. (2009). The first affordance originating from IoT-commerce are *context-aware services* that are enabled by sensors and actuators of the IoT device, as the environment can be observed and triggers for further action can be set based on sensor data. For instance, a smart thermostat might detect the need to heat a room and influence the amount of energy bought before the customer recognizes this need. As a second addition, IoT devices also allow n*atural interaction* with voice or gesture. For instance, voice assistants enable online shopping mainly controlled by voice commands. The third addition relates to intelligent algorithms in IoT devices that are used to *automate customer processes*. Decentralized intelligence embedded in IoT devices on the customer-side primarily emerged with IoT devices. Recommender systems and shopping agents that came up with e-commerce also inhibit intelligence but are usually situated on the supplier-side, whereas smart algorithms in IoT devices, enabling the actual automation of processes, are anchored on the customer-side. IoT devices, such as smart fridges, frequently lever this intelligence to decide autonomously when to re-order groceries.

The other affordances still remain valid, such as *information transparency*, which developed with e-commerce. It should be pointed out, however, that this also means that associated biases may still remain valid. Besides the positive effect of increased information transparency in e-commerce and m-commerce, research shows the existence of information bias. For instance, readers of online product reviews may be effected by sequential bias (Wan, 2015) and self-selection bias (Li & Hitt, 2008). The extent to which these biases still exist in IoT-commerce strongly depends on the specific setting with its inherent decisions (IoT device is autonomously deciding on a purchase vs. customer is deciding and purchasing via an IoT device). To check for the existence of biases in IoT-commerce is not the focus of this paper but is strongly recommended for future research.

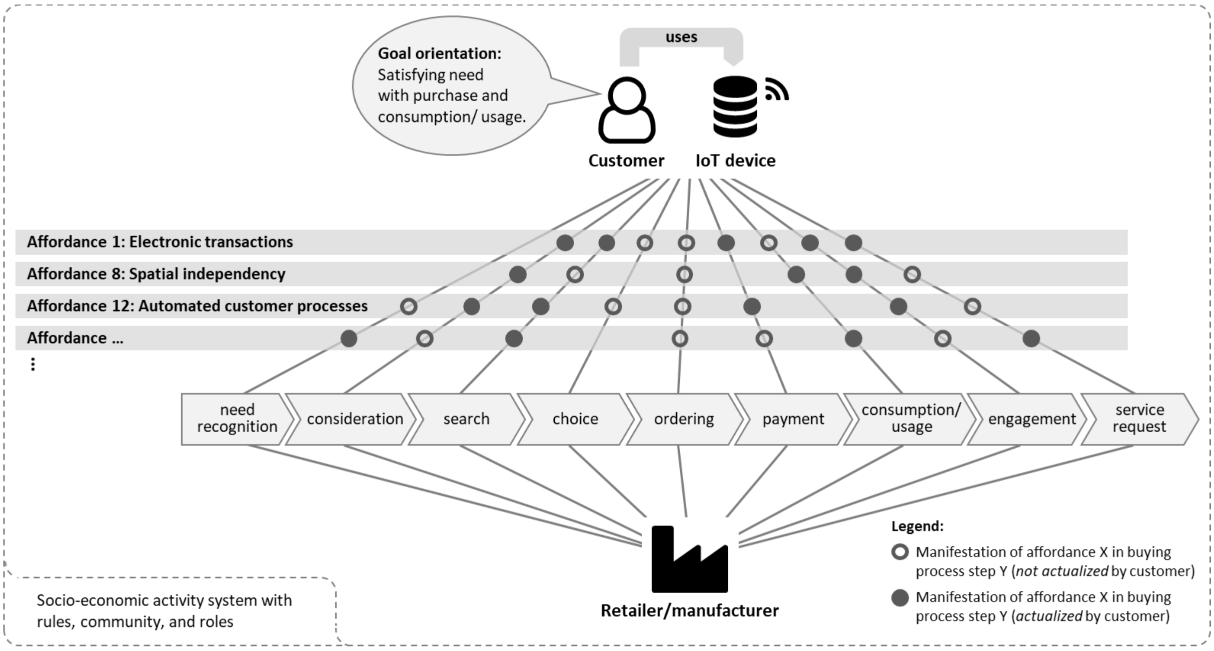

Goal orientation: Satisfying need with purchase and consumption/ usage.
uses
Customer
IoT device
Affordance 1: Electronic transactions
Affordance 8: Spatial independency
Affordance 12: Automated customer processes
Affordance ...
⋮
need recognition
consideration
search
choice
ordering
payment
consumption/ usage
engagement
service request
Retailer/manufacturer
Socio-economic activity system with rules, community, and roles
Legend:
○ Manifestation of affordance X in buying process step Y (not actualized by customer)
● Manifestation of affordance X in buying process step Y (actualized by customer)

Figure 2.1-2: Activity system of IoT-commerce with affordances and its manifestations

To visualize the role of affordances in the context of IoT-commerce, we depicted the related activity system schematically in Figure 2.1-2. Embedded in the socio-economic framework of rules (e.g., legislation), community (e.g., peer customers), and roles (e.g., social demography), the customer interacts with a retailer or manufacturer in different steps of the buying process. As described in the background section, we exemplarily relied on the 9-step buying process of Lemon and Verhoef (2016). In each step of the buying process, the customer (i.e., Person) might use her IoT device (i.e., Tool) in the interaction (i.e., goal-directed actions) with the retailer/manufacturer (i.e., Object) in order to satisfy her need with a purchase and the subsequent consumption/usage (i.e., goal orientation). Mainly derived from technical features and the socio-economic characteristics of IoT-commerce, the IoT device provides opportunities for action (i.e., affordances) to the customer. Each affordance (e.g., 'electronic transactions') might manifest in one or more steps of the buying process (e.g., 'ordering' and 'payment'). And in each step of the buying process (e.g., 'search'), one or more affordances (e.g., 'natural interaction' and 'automated customer processes') might manifest. These manifestations of affordances along the buying process might then be actualized if the customer performs respective goal-directed actions.

2.1.5 Validation for completeness and parsimony with real-life objects

We now validate the twelve affordances of IoT-commerce regarding completeness and parsimony. For this purpose, we assess the manifestations of the affordances along the customer buying process. This assessment levers the twelve affordances and a sample of 35 relevant IoT devices that were grouped into five categories such as ‘voice assistants’ and ‘replenishment services’. See Table 2.1-3 for definitions of the IoT device categories and exemplary product names. In this section, we show the relationships between all twelve affordances and the nine steps of the buying process. In Table 2.1-3, we use a bullet (‘●’) to indicate whether an affordance manifests in a step of the buying process. Table 2.1-3 shows the aggregated results whereas details can be found in Appendix A.

				Customer buying process								
				Pre-purchase stage			Purchase stage			Post-purchase stage		
				need recognition	consideration	search	choice	ordering	payment	consumption/ usage	engagement	service requests
IoT-commerce affordances	m-commerce affordance	e-commerce affordances	Electronic transactions		●	●	●	●	●	●	●	●
			Temporal independence		●	●		●	●	●	●	●
			Online platforms			●	●	●	●	●	●	●
			Information transparency		●	●	●			●		
			Social interactions	●	●	●	●	●	●	●	●	●
			Personalized services	●	●	●						●
			Proactive services	●	●	●						●
			Spatial independence		●	●		●		●	●	●
			Location-based services	●	●	●				●		●
			Context-aware services	●						●	●	●
			Natural interactions		●	●	●	●	●	●	●	●
			Automated customer processes	●	●	●	●	●	●		●	●

Legend: ● = Respective affordance (left column) manifests in the respective step of the customer buying process (top row).

Table 2.1-3: Manifestations of IoT-commerce affordances in the interaction between customer (i.e., Subject) and manufacturer/retailer/service provider (i.e., Object) via an IoT device (i.e., Tool)

Each step of the customer buying process comprises six to eleven manifestations of the affordances. *Social interactions* manifest in every step of the customer buying process, whereas *electronic transactions*, *natural interactions*, and *automated customer processes* lead to a manifestation in eight of nine steps. *Online platforms*, *temporal independence*, and *spatial independence* manifest in seven of nine steps. *Location-based services* (five of nine steps), *information transparency*, *personalized services*, *proactive services*, and *context-aware* *services* (four of nine steps) take effect in around half of all customer buying process steps. As presented in Table 2.1-3, all affordances manifest in multiple steps of the customer buying process. This suggests that the overview of IoT-commerce affordances is parsimonious. None of the affordances could be dropped without losing substantive content.

To check for completeness, two authors independently examined 337 IoT devices, filtered those relevant for retail commerce, and grouped them into five categories as described in Table 2.1-1. A check of these IoT devices did not yield any additional affordances not yet covered by Table 2.1-2, which was distilled based on extant academic literature. Therefore, we assume our overview of IoT-commerce affordances (Table 2.1-2) to be complete as all revealed opportunities to lever IoT devices in the context of retail commerce were already covered. Furthermore, the examination confirmed the manifestation of all affordances in at least four steps along the customer buying process as already observable today.

We see manifestations of affordances that are unique to IoT-commerce (e.g., using voice, haptics, gesture, or other natural interaction to search for products and services), and affordances that became apparent with e-commerce or m-commerce and now continue within IoT-commerce (e.g., perceiving a lowered barrier to pursue the fulfillment of a need) – perhaps even intensified due to comfort that comes along with IoT devices. Note that the steps of the buying process and consequently the manifestations of the affordances are not necessarily in the mentioned order, but can vary (e.g., a customer can engage with an organization before the actual consumption, for instance when writing a product review about the ordering process before receiving/using the product). Furthermore, some steps can be skipped (e.g., a customer spontaneously considers buying a product without searching for alternatives, directly jumping from ‘consideration’ to ‘choice’). In the following, we provide details on the manifestations and real-life examples.

The steps of the pre-purchase stage (i.e., ‘need recognition’, ‘consideration’, and ‘search’) comprise six to eleven manifestations of affordances in each step. In the step ‘need recognition’, customers might be guided by *personalized* advertisements and *proactive* recommendations for products and services that may even be *dependent on the location* of the IoT device and/or customer. Customers may also recognize unidentified needs through *social interaction* with their peers via social networks and the encountered content. With its *context-aware* sensors, IoT devices offer an additional alternative to recognize the needs of customers. For instance, solutions for smart resource management such as smart thermostats or smart sprinklers collect environmental data like room temperature or weather conditions. With smart algorithms, the

IoT device analyzes this sensor data to evaluate whether the room temperature is too cold or the lawn demands watering, without the customer triggering this process of automated need recognition (*automated customer process*). This is a brief example of the manifestation of IoT affordances in the first step of the customer buying process, namely ‘need recognition’. Further exemplifications for the eight following steps can be found in Appendix B.

2.1.6 Discussion

From a theoretical perspective, we examined extant literature and research streams in the field of retail commerce and structured it along three waves, namely e-commerce, m-commerce, and IoT-commerce. We also investigated the literature of IoT that so far primarily focused on technical and business-related but not on customer-focused aspects. Bringing together both domains, we shed light and extend the body of knowledge at the intersection of retail commerce and IoT that we call IoT-commerce. We investigate this field with a customer-centric IS perspective. Conceptualizing IoT-commerce is the first theoretical contribution of this paper. In particular, we analyzed the influence and opportunities of IoT devices in the customer buying process. This is especially relevant due to its fundamental impact on both customers and companies. In this, we identified twelve affordances of IoT-commerce from a customer perspective. The identification, conceptualization, and linkage of these affordances to the customer buying process is this paper‘s second theoretical contribution.

Nine of the twelve affordances of IoT-commerce are already known from e-commerce and m-commerce. It is important to note that they carry on in IoT-commerce. Within IoT-commerce, they might be present or might be actualized more frequently than before. However, these nine affordances are not qualitatively new and, thus, less disruptive than the new affordances. Three affordances of IoT-commerce – namely context-aware services, natural interactions, and automated customer processes – are qualitatively new as compared to prior forms of IT-enabled retail commerce. As our analysis highlights, these three new affordances jointly affect each step in the customer buying process. Current real-life examples of IoT devices already demonstrate how these affordances manifest along the buying process. However, we are only at the beginning of the IoT era. On the one hand, IoT devices are about to spread into private homes and lives transforming online shopping at a fast pace. On the other hand, organizations constantly enlarge functionalities of IoT devices in order to gather more data of individual users, provide more convenience and service, and better predict individual user behavior. As our affordances and their respective manifestations show, IoT has the potential to innovate the customer buying process we currently know from e-commerce and m-commerce.

Consequently, we are convinced that IoT-commerce is a highly relevant research topic that is gaining considerably in importance within the next years. With our research, we contribute to its theoretical foundation and offer insight into IoT-commerce from a customer’s point of view.

Our work itself is theoretically founded in Activity Theory and Affordance Theory. A combination of both was a suitable tool to develop our theory of IoT-commerce based on extant theory in the areas of e-commerce, m-commerce, and IoT. In Table 2.1-4, we briefly summarize our theory components and evaluate them. Gregor (2006) presented a widely used typology of theories in IS research (about 3,000 citations according to Google Scholar). Based on four primary goals of theory (analysis and description, explanation, prediction, prescription) she identified different types of theories and components of such theories (Table 2.1-3 of Gregor, 2006). Within these components, four components are common to all theories: means of representation, constructs, statements of relationship, and scope. In Table 2.1-4, we use these four mandatory components of theories in IS as a structure to present our theory of IoT-commerce as we believe that following this structure adds clarity to the presentation of our contribution.

To summarize the evaluation of our theory, we refer to the criteria suggested by Weber (2012). Weber presents a detailed framework and criteria for evaluating theories in IS research. He presents criteria relating to the different parts of a theory individually and to the theory as a whole. For brevity of presentation, we restrict the discussion in Table 2.1-4 to the five criteria for the theory as a whole (importance, novelty, parsimony, level, falsifiability) as these appear to us more insightful for the theory at hand than the criteria for individual parts.

Theory Component (as proposed by Gregor, 2006)	
Means of representation	Our theory of IoT-commerce is described in words, tables, and pictures. Words are used for detailed explanations enriched by examples. Tables are used to structure the main constructs of our theory, the *affordances* of IoT-commerce as well as the manifestations of those *affordances* along the buying process. Schematic pictures illustrate the main constructs within the *activity system* as a theoretical foundation.
Constructs	Constructs comprise the customer (i.e., *Person* and goal-oriented actor), the IoT device (i.e., *Tool*), the retailer/manufacturer (i.e., *Object*), steps in the customer buying process (i.e., nine steps in three stages), three waves of commerce evolution (i.e., e-commerce, m-commerce, and IoT-commerce), and the generic concept of *affordances* in context of retail commerce (i.e., opportunities for goal-directed actions). In particular, we present twelve *affordances* of IoT-commerce (e.g., natural interaction) and their manifestations along the buying process (e.g., using voice commands in the step of 'search'). The constructs used in our theory are itself theoretically founded in *Activity Theory* and *Affordance Theory*.
Statements of relationship	The relationship between the customer, the IoT device, and the retailer/manufacturer is derived from *Activity and Affordance Theories* and described in detail. Based on this theoretical foundation, we explain the relationship of the twelve *affordances* with the three waves of retail commerce. We furthermore present relationships between *affordances* and the steps of the buying process in the sense of manifestations.
Scope	As the majority of the examined literature is composed by European, American, and Asian researchers, our theory shall be applicable in those regions. However, as we expect technological development to continuously spread further, we are convinced that our theory holds true for nearly all geographic regions and social demographics. Importantly, it is restricted to retail commerce and does not cover business-to-business (B2B) commerce.
Evaluation Criterion (as proposed by Weber, 2012)	
Importance	Our theory provides insights into the changing nature of retail commerce driven by the diffusion of the IoT. Impacting customer behavior, creating technological opportunities, and potentially facilitating innovative business models, IoT-commerce should be considered as an important domain for both researchers and practitioners. Our theory conceptualizes IoT-commerce and identifies its *affordances*. As such, it is a basis for further research in this area of growing practical relevance.
Novelty	Driven by the emerging phenomenon of IoT, our theory about IoT-commerce provides insights into the evolution of retail commerce that no researcher has examined in detail yet.
Parsimony	Our theory comprises a conceivable small number of constructs and omits aspects not directly relevant for the explanatory power of the theory (such as legislative aspects). *Affordances* and their manifestations are presented compactly.
Level	Our research is framed as a middle-range (meso) theory avoiding 'narrow empiricism' and 'over-generalization'.
Falsifiability	With the transformation of *affordances* into manifestations along the buying process, our theory can be tested if all affordances actually manifest and potentially be falsified if different observations are made.

Table 2.1-4: Components and evaluation of theory

From a practical point of view, our paper on the opportunities of IoT devices serves as a tool for customers as well as companies. For retail customers, it initiates critical reflections about the use of IoT devices in the buying process. We provide them with insights on how IoT contributes to their customer experience, for instance, that customer activity is less necessary and that the barrier to pursue the fulfillment of a need decreases significantly. Furthermore, customers comprehend better when IoT decreases their sovereignty and self-determination, initiating reflections about the trend towards automated decision-making with a high volume of data collection. For organizations, we provide a theoretical foundation and structure to analyze their products and services in order to identify opportunities (e.g., enhance customer experience, increase customer loyalty, create lock-in effects) and risks (e.g., speed of IoT usage by competitors, slow adaption to changing customer behavior). For organizations, it is important to analyze the impact of IoT-commerce on their business model and initiate strategies to minimize risks and increase opportunities. For instance, IoT bears the opportunity for manufacturers to sell directly to the customer, and therefore establish a direct customer relationship by omitting a retailer. In contrast, retailers risk to lose direct customer contact in case manufactures directly sell to end customers. Those aspects must be assessed and considered in suitable business models. Overall, we found the most interesting manifestations of affordances newly arising from IoT devices in the pre- and post-purchase stages, whereas the purchase stage is primarily characterized by a reinforcement of the affordances that already emerged with e-commerce and m-commerce as well as natural interaction of the customer with the device and automation of choice, ordering, and payment.

As any research endeavor, our paper is beset with limitations that stimulate further research. First, we do not compile a full set of all existing (or even future) IoT devices. Although we drew on a broad sample from three publications with 337 devices in total, filtered to 35 relevant examples, and generalized into five categories, innovative IoT devices will emerge and may afford new actions for future retail customers. Of course, we cannot assure that we covered all kinds of devices that might come up. Further research could consider upcoming IoT devices in the next years and, if necessary, revise our IoT device categories, redo the validation for completeness and parsimony, and extend our overview of IoT-commerce affordances if completely new functions emerged. Furthermore, our paper focuses on the retail commerce context (B2C). Given the numerous usages of IoT devices in the B2B context, further research should check the applicability of our affordances in B2B commerce.

Future research might want to go beyond the mitigation of the above limitations and investigate into follow-up questions such as: Which differences can be observed between real and perceived affordances (e.g., role of tech-savviness or cultural background of customers)? How do customers accept the new technology-driven phenomenon of IoT-commerce (e.g., are these conscious perception and actualization processes)? How do the affordances change customer purchasing behavior over time and compared to e-commerce and m-commerce (e.g. changes in brand loyalty, frequency of purchases, or in the relative weight of different purchasing criteria)? Are there any negative side effects for customers associated with IoT-commerce (e.g., self-creation of lock-in effects)? How can established and emerging companies lever the opportunities of IoT-commerce to offer and monetize additional customer value (e.g., how to integrate IoT-commerce into multi- or omnichannel customer interaction)? To which biases (e.g., information bias) and nudges (e.g., recommendations) are customers exposed in IoT-commerce and how can businesses lever IoT-spawned technology to overcome or utilize these biases and nudges? Other work beyond IoT-commerce affordances might want to focus on the role of legislation, the influence of data privacy, a rigorously developed taxonomy of IoT devices in retail commerce, emerging business models, or an ethical perspective (e.g., potential moral issues affiliated with functionality such as automated decision-making).

2.1.7 Conclusion

This work was motivated by analyzing IoT in the context of commerce in a customer-centric manner. Through an affordance lens, we answered our research question which opportunities IoT devices provide to retail commerce customers. Theoretically founded in Activity Theory and Affordance Theory, we develop our theory of IoT-commerce as a third wave in the evolution of retail commerce following e-commerce and m-commerce. We identified 49 relevant articles in a structured literature search in leading IS journals, commerce journals, marketing journals, and the domain of computer science and therefrom extracted twelve affordances. Seven affordances emerged with e-commerce, two with m-commerce, and three additional affordances originate from IoT-commerce. To evaluate our theory and to demonstrate its applicability, we derived manifestations of the twelve affordances along with the nine steps in three stages of the customer buying process. We further extracted 35 relevant IoT devices out of a sample with 337 real-life IoT devices, grouped them into five categories, and levered them to confirm completeness and parsimony. Overall, our paper helps understand the influence of the IoT phenomenon on retail commerce in a customer-centric manner.

As IoT-commerce is still in its infancy, its future holds tremendous potential. Due to the increasing proliferation of IoT devices, its importance for retail commerce may continuously rise. Similar to the trend of customers preferring mobile devices before desktop PCs for ordering online, IoT devices might become a vital – or even the most vital – channel for retail commerce in the near future. Hence, detailed analysis and understanding of the IoT-commerce phenomenon and its consequences for both customers and companies is of utmost importance and represents the crucial basis for further goal-directed action. Academic research might, therefore, want to keep up with this fast-evolving phenomenon and answer related questions such as those raised in the previous section.

References

Adolphs, C., & Winkelmann, A. (2010). Personalization research: a rigorous literature review on personalization in e-commerce (2000-2008). *Journal of Electronic Commerce Research*, 11(4), 326–341.

AIS. (2018). Senior scholars' basket of journals. https://aisnet.org/page/SeniorScholarBasket

Akter, S., & Wamba, S. F. (2016). Big data analytics in e-commerce: a systematic review and agenda for future research. *Electronic Markets*, 26(2), 173–194. https://doi.org/10.1007/s12525-016-0219-0

Aldridge, D. (1998). Purchasing on the net: the new opportunities for electronic commerce. *Electronic Markets*, 8(1), 34–37. https://doi.org/10.1080/10196789800000010

Allen, D. K., Brown, A., Karanasios, S., & Norman, A. (2013). How should technology-mediated organizational change be explained? A comparison of the contributions of critical realism and activity theory. *MIS Quarterly*, 37(3), 835–854.

Al-Obeidat, F., Spencer, B., & Kafeza, E. (2018). The opinion management framework: identifying and addressing customer concerns extracted from online product reviews. *Electronic Commerce Research and Applications*, 27, 52–64. https://doi.org/10.1016/j.elerap.2017.11.003

Alt, R., & Klein, S. (2011). Twenty years of electronic markets research: looking backwards towards the future. *Electronic Markets*, 21(1), 41–51. https://doi.org/10.1007/s12525-011-0057-z

Amazon. (2018a). *Amazon echo*. https://www.amazon.com/echo

Amazon. (2018b). *Dash replenishment service*. https://developer.amazon.com/de/dash-replenishment-service

Baek, H., Ahn, J., & Choi, Y. (2012). Helpfulness of online consumer reviews: readers' objectives and review cues. *International Journal of Electronic Commerce*, 17(2), 99–126. https://doi.org/10.2753/JEC1086-4415170204

Bærentsen, K. B., & Trettvik, J. (2002). An activity theory approach to affordance. In *Proceedings of the 2nd nordic conference on human-computer interaction* (nordichi 2002), Aarhus, Denmark.

Baethge, C., Klier, J., & Klier, M. (2016). Social commerce: state-of-the-art and future research directions. *Electronic Markets*, 26(3), 269–290.

Barbosa, J. L. V. (2015). Ubiquitous computing: Applications and research opportunities. In *Proceedings of the ieee international conference on computational intelligence and computing research* (iccic 2015), Madurai, India.

Beaudry, A., & Carillo, K. D. (2006). The customer-centered b2c literature through the lens of activity theory: a review and research agenda. *Communications of the Association for Information Systems*, 17(1), 428-503.

Benbunan-Fich, R. (2019). An affordance lens for wearable information systems. *European Journal of Information Systems*, 28(3), 256-271.

Bhajantri, L. B., Nalini, N., & Rathod, S. H. (2015). Collaborative filtering technique based recommendation in ubiquitous commerce. In *Proceedings of the international conference on applied and theoretical computing and communication technology* (icatcct 2015), Davangere, India.

Borgia, E. (2014). The internet of things vision: key features, applications and open issues. *Computer Communications*, 54, 1–31. https://doi.org/10.1016/j.comcom.2014.09.008

Brandt, R., Püschel, L., Röglinger, M., & Schlott, H. (2017). Unravelling the internet of things: a multi-layer taxonomy and archetypes of smart things. *Working Paper of the FIM Research Center.*

Brynjolfsson, E., & Smith, M. D. (2000). Frictionless commerce? A comparison of internet and conventional retailers. *Management Science*, 46(4), 563–585.

Butler, P., & Peppard, J. (1998). Consumer purchasing on the internet: processes and prospects. *European Management Journal*, 16(5), 600–610.

Cambridge Dictionary. (2019a). *Definition of "product".* https://dictionary.cambridge.org/de/worterbuch/englisch/product

Cambridge Dictionary. (2019b). *Definition of "service".* https://dictionary.cambridge.org/de/worterbuch/englisch/service

Chaudhury, A., Mallick, D. N., & Rao, H. R. (2001). Web channels in e-commerce. *Communications of the ACM*, 44(1), 99–104.

Chua, C., Khoo, H., Straub, D., Kadiyala, S., & Kuechler, D. (2005). The evolution of e-commerce research: a stakeholder perspective. *Journal of Electronic Commerce Research*, 6(4), 262–280.

Chunxia, Q., Zhenzhong, Z., & Litao, Z. (2010). A shopping model in ubiquitous media environment. In *Proceedings of the international forum on information technology and applications* (ifita 2010), Kunming, China.

Clarke, I. I. (2008). Emerging value propositions for m-commerce. *Journal of Business Strategies*, 25(2), 41–57.

Coughlan, T., Brown, M., Mortier, R., Houghton, R. J., Goulden, M., & Lawson, G. (2012). Exploring acceptance and consequences of the internet of things in the home. In *Proceedings of the ieee international conference on green computing and communications* (greencom 2012), Besancon, France.

Cui, Y., Mou, J., & Liu, Y. (2018). Knowledge mapping of social commerce research: a visual analysis using citespace. *Electronic Commerce Research*, 18(4), 837–868. https://doi.org/10.1007/s10660-018-9288-9

Cunningham, L. F., Gerlach, J. H., Harper, M. D., & Young, C. E. (2005). Perceived risk and the consumer buying process: internet airline reservations. *International Journal of Service Industry Management*, 16(4), 357–372. https://doi.org/10.1108/09564230510614004

Deloitte. (2016). *Switch on to the connected home: The Deloitte Consumer Review*. https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consumer-business/deloitte-uk-consumer-review-16.pdf

Engel, J. F., Blackwell, R. D., & Miniard, P. W. (1995). *Consumer behavior* (8. ed.). Dryden.

Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Orienta-Konsultit Oy.

Evans, M. (2017, January 24). *5 ways the internet of things will influence commerce*. Forbes. https://www.forbes.com/sites/michelleevans1/2017/01/24/5-ways-the-internet-of-things-will-influence-commerce

Evans, M. (2018, May 31). *Iot will have the most impact on business in the next five years, survey says*. Forbes. https://www.forbes.com/sites/michelleevans1/2018/05/31/iot-will-have-the-most-impact-on-business-in-the-next-five-years-survey-says

Farhad, M. (2018, October 10). *A future where everything becomes a computer is as creepy as you feared*. New York Times. https://www.nytimes.com/2018/10/10/technology/future-internet-of-things.html

Fleisch, E., Sarma, S., & Thiesse, F. (2009). Preface to the focus theme section: ‘internet of things’. *Electronic Markets*, 19(2-3), 99–102. https://doi.org/10.1007/s12525-009-0016-0

Fox, P., Rezania, D., Wareham, J., & Christiaanse, E. (2006). Will mobiles dream of electric sheep? Expectations of the new generation of mobile users: misfits with practice and research. In *Proceedings of the international conference on mobile business* (icmb 2006).

Frambach, R. T., Roest, H. C. A., & Krishnan, T. V. (2007). The impact of consumer internet experience on channel preference and usage intentions across the different stages of the buying process. *Journal of Interactive Marketing*, 21(2), 26–41. https://doi.org/10.1002/dir.20079

Fu, D., Hong, Y., Wang, K., & Fan, W. (2018). Effects of membership tier on user content generation behaviors: evidence from online reviews. *Electronic Commerce Research*, 18(3), 457–483.

Gaskin, J., Berente, N., Lyytinen, K., & Yoo, Y. (2014). Toward generalizable sociomaterial inquiry: a computational approach for zooming in and out of sociomaterial routines. *MIS Quarterly*, 38(3), 849–871. https://doi.org/10.25300/MISQ/2014/38.3.10

Gaur, A., Scotney, B., Parr, G., & McClean, S. (2015). Smart city architecture and its applications based on iot. *Procedia Computer Science*, 52, 1089–1094.

Gaver, W. W. (1991). Technology affordances. In *Proceedings of the sigchi conference on human factors in computing systems* (chi 1991), New Orleans, Louisiana, USA.

Gaver, W. W. (1992). The affordances of media spaces for collaboration. In *Proceedings of the acm conference on computer-supported cooperative work* (cscw 1992), Toronto, ON, Canada.

Gefen, D., Karahanna, E., & Straub, D. (2003). Trust and tam in online shopping: an integrated model. *MIS Quarterly*, 27(1), 51–90.

Gengatharen, D. E., & Standing, C. (2005). A framework to assess the factors affecting success or failure of the implementation of government-supported regional e-marketplaces for smes. *European Journal of Information Systems*, 14(4), 417–433. https://doi.org/10.1057/palgrave.ejis.3000551

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Grandon, E. E., & Pearson, J. M. (2004). Electronic commerce adoption: an empirical study of small and medium US businesses. *Information & Management*, 42(1), 197–216.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, 30(3), 611–642.

Groenfeldt, T. (2016, January 6). *Touchscreen on the fridge: door to order groceries, watch football.* Forbes, 2016. https://www.forbes.com/sites/tomgroenfeldt/2016/01/06/touchscreen-on-the-fridge-door-to-order-groceries-watch-football

Groß, M. (2015). Mobile shopping: a classification framework and literature review. *International Journal of Retail & Distribution Management*, 43(3), 221–241. https://doi.org/10.1108/IJRDM-06-2013-0119

Guttman, R., Moukas, A., & Maes, P. (1999). *Agents as mediators in electronic commerce*. In M. Klusch (Ed.), Intelligent information agents: agent-based information discovery and management on the internet (pp. 131–152). Springer. https://doi.org/10.1007/978-3-642-60018-0_8

Hartson, H. R. (2003). Cognitive, physical, sensory, and functional affordances in interaction design. *Behavior & Information Technology*, 22(5), 315–338.

Hasan, H., & Kazlauskas, A. (2014). *Activity theory: who is doing what, why and how*. In H. Hasan (Ed.), Being practical with theory: a window into business research (pp. 9–14). THEORI.

Heatman, A. (2018). *Fridgecam is the internet of things device that will actualy improve your life.* https://www.standard.co.uk/tech/fridgecam-internet-of-things-food-waste-a3891031.html

Ho, Y.-C., Wu, J., & Tan, Y. (2017). Disconfirmation effect on online rating behavior: a structural model. *Information Systems Research*, 28(3), 626–642. https://doi.org/10.1287/isre.2017.0694

Hoffman, D. L., & Novak, T. P. (2018). Consumer and object experience in the internet of things: an assemblage theory approach. *Journal of Consumer Research*, 44(6), 1178–1204.

Hollocks, B. W. (2001). Book review: handbook on electronic commerce. European Journal of Information Systems, 10(1), 69. https://doi.org/10.1046/j.1365-3113.2001.00158.x

Howard, J. A., & Sheth, J. N. (1969). *The theory of buyer behavior*. The Wiley marketing series. Wiley.

Huang, Z., & Benyoucef, M. (2013). From e-commerce to social commerce: A close look at design features. *Electronic Commerce Research and Applications*, 12(4), 246–259. https://doi.org/10.1016/j.elerap.2012.12.003

Huber, R., Püschel, L., & Röglinger, M. (2017). Iot-enabled smart service systems: identification of actors and interaction types. *Working Paper of the FIM Research Center*.

Jing, N., Jiang, T., Du, J., & Sugumaran, V. (2018). Personalized recommendation based on customer preference mining and sentiment assessment from a chinese e-commerce website. *Electronic Commerce Research*, 18(1), 159–179. https://doi.org/10.1007/s10660-017-9275-6

Johnston, R. B., & Gregor, S. (2000). A theory of industry-level activity for understanding the adoption of interorganizational systems. *European Journal of Information Systems*, 9(4), 243–251.

Junglas, I., & Watson, R. (2003). U-commerce: a conceptual extension of e-commerce and m-commerce. In *Proceedings of the international conference on information systems* (icis 2003) (Article 55).

Kaptelinin, V. (2005). *Activity theory*. In encyclopaedia of human computer interaction. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/activity-theory

Kaptelinin, V., & Nardi, B. (2012). Affordances in hci: toward a mediated action perspective. In *Proceedings of the sigchi conference on human factors in computing systems* (chi 2012), Austin, Texas, USA.

Karanasios, S., & Allen, D. K. (2014). Mobile technology in mobile work: contradictions and congruencies in activity systems. *European Journal of Information Systems*, 23(5), 529–542.

Kees, A., Oberländer, A. M., Röglinger, M., & Rosemann, M. (2015). Understanding the internet of things: a conceptualisation of business-to-thing (b2t) interactions. In *Proceedings of the 23th European conference on information systems* (ecis 2015), Münster, Germany.

Klopping, I. M., & McKinney, E. (2004). Extending the technology acceptance model and the task-technology fit model to consumer e-commerce. *Information Technology, Learning, and Performance Journal*, 22(1), 35–47.

Kocas, C. (2002). Evolution of prices in electronic markets under diffusion of price-comparison shopping. *Journal of Management Information Systems*, 19(3), 99–119.

Konjengbam, A., Dewangan, N., Kumar, N., & Singh, M. (2018). Aspect ontology based review exploration. *Electronic Commerce Research and Applications*, 30, 62–71. https://doi.org/10.1016/j.elerap.2018.05.006

Koverman, C. (2016). Next-generation connected support in the age of iot: it's time to get proactive about customer support. *IEEE Consumer Electronics Magazine,* 5(1), 69–73.

Kozinets, R. V. (2019). Consuming technocultures: an extended jcr curation. Journal of Consumer Research, 46(3), 620–627.

Kraemer, K. L., Dedrick, J., & Yamashiro, S. (2000). Refining and extending the business model with information technology: Dell computer corporation. *The Information Society,* 16(1), 5–21.

Kurkovsky, S. (2005). Using principles of pervasive computing to design m-commerce applications. In *Proceedings of the international conference on information technology: Coding and computing* (itcc 2005), Las Vegas, Nevada, USA.

Laudon, K. C., & Traver, C. G. (2018). *E-commerce: Business, technology, society* (13th ed.). Pearson.

Lee, H. G., & Lee, R. M. (1993). Intelligent electronic trading for commodity exchanges. *Electronic Markets,* 9-10, 5–6.

Lee, I., & Lee, K. (2015). The internet of things (iot): applications, investments, and challenges for enterprises. *Business Horizons,* 58(4), 431–440.

Lee, S. M., Hwang, T., & Lee, D. H. (2011). Evolution of research areas, themes, and methods in electronic commerce. *Journal of Organizational Computing and Electronic Commerce,* 21(3), 177–201. https://doi.org/10.1080/10919392.2011.590095

Lee, Y. E., & Benbasat, I. (2004). A framework for the study of customer interface design for mobile commerce. *International Journal of Electronic Commerce,* 8(3), 79–102.

Lehrer, C., Wieneke, A., Vom Brocke, J., Jung, R., & Seidel, S. (2018). How big data analytics enables service innovation: materiality, affordance, and the individualization of service. *Journal of Management Information Systems,* 35(2), 424–460. https://doi.org/10.1080/07421222.2018.1451953

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing,* 80(6), 69–96. https://doi.org/10.1509/jm.15.0420

Leonardi, P. M. (2011). When flexible routines meet flexible technologies: affordance, constraint, and the imbrication of human and material agencies. *MIS Quarterly,* 35(1), 147–167.

Leong, C., Pan, S. L., Newell, S., & Cui, L. (2016). The emergence of self-organizing e-commerce ecosystems in remote villages of china: a tale of digital empowerment for rural development. *MIS Quarterly*, 40(2), 475–484. https://doi.org/10.25300/MISQ/2016/40.2.11

Leont'ev, A. N. (1978). *Activity, Consciousness, and Personality*. Prentice-Hall.

Levina, O., & Vilnai-Yavetz, I. (2015). E-visibility maturity model: A tool for assessment and comparison of individual firms and sets of firms in e-business. *Electronic Commerce Research and Applications*, 14(6), 480–498. https://doi.org/10.1016/j.elerap.2015.07.004

Li, S., & Karahanna, E. (2015). Online recommendation systems in a b2c e-commerce context: a review and future directions. *Journal of the Association for Information Systems*, 16(2), 72–107. https://doi.org/10.17705/1jais.00389

Li, X [Xinxin], & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 19(4), 456–474.

Li, X [Xuemei], Xu, G [Gang], & Li, L. (2008). Rfid based smart home architecture for improving lives. In *Proceedings of the 2nd ieee international workshop on anti-counterfeiting, security, identification* (asid 2008), Guiyang, China.

Liang, T.-P., Ho, Y.-T., Li, Y.-W., & Turban, E. (2011). What drives social commerce: the role of social support and relationship quality. *International Journal of Electronic Commerce*, 16(2), 69–90. https://doi.org/10.2753/JEC1086-4415160204

Liu, Q., Huang, S., & Zhang, L. (2016). The influence of information cascades on online purchase behaviors of search and experience products. *Electronic Commerce Research*, 16(4), 553–580. https://doi.org/10.1007/s10660-016-9220-0

Lynch, J. G., & Ariely, D. (2000). Wine online: search costs affect competition on price, quality, and distribution. *Marketing Science*, 19(1), 83–103.

Ma, Y., Chen, G., & Wei, Q. (2017). Finding users preferences from large-scale online reviews for personalized recommendation. *Electronic Commerce Research*, 17(1), 3–29. https://doi.org/10.1007/s10660-016-9240-9

Maity, M., & Dass, M. (2014). Consumer decision-making across modern and traditional channels: E-commerce, m-commerce, in-store. *Decision Support Systems*, 61, 34–46. https://doi.org/10.1016/j.dss.2014.01.008

Manvi, S. S., Nalini, N., & Bhajantri, L. B. (2011). Recommender system in ubiquitous commerce. In *Proceedings of the 3rd international conference on electronic computer technology*, Kanyakumari, India.

Markus, M. L., & Silver, M. S. (2008). A foundation for the study of it effects: a new look at desanctis and poole’s concepts of structural features and spirit. *Journal of the Association for Information Systems,* 9(10/11), 609–632.

McGrenere, J., & Ho, W. (2000). Affordances: clarifying and evolving a concept. In *Proceedings of graphics interface* (gi 2000), Montréal, Québec, Canada.

McKinsey (2013). *Disruptive technologies: Advances that will transform life, business, and the global economy.* https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/disruptive-technologies

Mengxiang, L., Chuan-Hoo, T., Kwok-Kee, W., & Kanliang, W. (2017). Sequentiality of product review information provision: an information foraging perspective. *MIS Quarterly,* 41(3), 867–892.

Miles, I. (1990). Teleshopping: just around the corner? *RSA Journal,* 138(5403), 180–189.

Nan, G., Yang, J., & Dou, R. (2017). Do only review characteristics affect consumers' online behaviours? A study of relationship between reviews. *Journal of Electronic Commerce Research,* 18(4), 330-345.

Nardi, B. (1996). *Context and consciousness: Activity theory and human–computer interaction*. MIT Press.

Nassauer, S. (2017, May 4). Wal-mart wants to know when your milk is about to expire. *Wall Street Journal.* https://www.wsj.com/articles/wal-mart-wants-to-know-when-your-milk-is-about-to-expire-1493937138

Ngai, E. W. T., & Gunasekaran, A. (2007). A review for mobile commerce research and applications. *Decision Support Systems,* 43(1), 3–15. https://doi.org/10.1016/j.dss.2005.05.003

Ngai, E. W. T., & Wat, F. K. T. (2002). A literature review and classification of electronic commerce research. *Information & Management,* 39(5), 415–429. https://doi.org/10.1016/S0378-7206(01)00107-0

Nicosia, F. M., & Mayer, R. N. (1976). Toward a sociology of consumption. *Journal of Consumer Research,* 3(2), 65–75.

Norman, D. A. (1988). *The psychology of everyday things*. Basic Books.

Norman, D. A. (1999). Affordance, conventions, and design. *Interactions,* 6(3), 38–43.

Oberländer, A. M., Röglinger, M., Rosemann, M., & Kees, A. (2018). Conceptualizing business-to-thing interactions: a sociomaterial perspective on the internet of things. *European Journal of Information Systems*, 27(4), 486–502.

Oxford Dictionary. (2018). *Commerce*. https://en.oxforddictionaries.com/definition/commerce

Peng, L., Liao, Q., Wang, X., & He, X. (2016). Factors affecting female user information adoption: an empirical investigation on fashion shopping guide websites. *Electronic Commerce Research*, 16(2), 145–169. https://doi.org/10.1007/s10660-016-9213-z

Porter, M. E., & Heppelmann, J. E. (2014). How smart, connected products are transforming companies. *Harvard Business Review*. https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition

Pousttchi, K., Tilson, D., Lyytinen, K., & Hufenbach, Y. (2015). Introduction to the special issue on mobile commerce: mobile commerce research yesterday, today, tomorrow - what remains to be done? *International Journal of Electronic Commerce,* 19(4), 1–20. https://doi.org/10.1080/10864415.2015.1029351

Prasad, B. (2003). Intelligent techniques for e-commerce. *Journal of Electronic Commerce Research*, 4(2), 65–71.

Püschel, L., Schlott, H., & Röglinger, M. (2016). What's in a smart thing? Development of a multi-layer taxonomy. In *Proceedings of the international conference on information systems* (icis 2016), Dublin, Ireland.

Ramaswamy, V., & Ozcan, K. (2018). Offerings as digitalized interactive platforms: a conceptual framework and implications. *Journal of Marketing*, 82(4), 19–31.

Rao, B., & Minakakis, L. (2003). Evolution of mobile kocation-based services. *Communications of the ACM*, 46(12), 61–65.

Reid, T. (2018, December 19). *Everything alexa learned in 2018*. https://blog.aboutamazon.com/devices/everything-alexa-learned-in-2018

Romano, N. C., JR., & Fjermestad, J. (2002). Electronic commerce customer relationship management: an assessment of research. *International Journal of Electronic Commerce,* 6(2), 61–113.

Rosemann, M. (2014). Proposals for future bpm research directions. In *Proceedings of the 2nd asia pacific business process management conference.*

Rothensee, M. (2008). User acceptance of the intelligent fridge: empirical results from a simulation. In *Proceedings of the 1st international conference on the internet of things* (iot 2008) (pp. 123–139).

Safi, R., & Yu, Y. (2017). Online product review as an indicator of users' degree of innovativeness and product adoption time: a longitudinal analysis of text reviews. *European Journal of Information Systems*, 26(4), 414–431.

Samaras, G. (2002). Mobile commerce: vision and challenges (location and its management). In *Proceedings of the symposium on applications and the internet*.

Sanchez-Pi, N., & Molina, J. M. (2009). A multi-agent platform for the provisioning of U-commerce services. NAFIPS 2009-2009 *Annual Meeting of the North American Fuzzy Information Processing Society*.

Saumya, S., Singh, J. P., Baabdullah, A. M., Rana, N. P., & Dwivedi, Y. K. (2018). Ranking online consumer reviews. *Electronic Commerce Research and Applications*, 29, 78–89. https://doi.org/10.1016/j.elerap.2018.03.008

Schierz, P. G., Schilke, O., & Wirtz, B. W. (2010). Understanding consumer acceptance of mobile payment services: An empirical analysis. *Electronic Commerce Research and Applications*, 9(3), 209–216. https://doi.org/10.1016/j.elerap.2009.07.005

Shang, X., Zhang, R., & Chen, Y. (2012). Internet of things (iot) service architecture and its application in e-commerce. *Journal of Electronic Commerce in Organizations*, 10(3), 44–55. https://doi.org/10.4018/jeco.2012070104

Sharma, S., & Gutiérrez, J. A. (2010). An evaluation framework for viable business models for m-commerce in the information technology sector. *Electronic Markets*, 20(1), 33–52. https://doi.org/10.1007/s12525-010-0028-9

Shim, J., Avital, M., Dennis, A. R., Rossi, M., Sørensen, C., & French, A. (2019). The transformative effect of the internet of things on business and society. *Communications of the Association for Information Systems*, 44(1), 129-140.

Song, Z., Sun, Y., Wan, J., Huang, L., & Zhu, J. (2017). Smart e-commerce systems: current status and research challenges. *Electronic Markets*, 26(3), 473. https://doi.org/10.1007/s12525-017-0272-3

Statista. (2018a). *Digital shopping device usage and frequency worldwide in 2017*. https://www.statista.com/statistics/692846/online-shopping-device-worldwide-frequency/

Statista. (2018b). *Retail e-commerce sales worldwide from 2014 to 2021 (in billion U.S. Dollars).* https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

Strader, T. J., & Shar, M. J. (1997). Characteristics of electronic markets. *Decision Support Systems,* 21(3), 185–198.

Timmers, P. (1998). Business models for electronic markets. *Electronic Markets*, 8(2), 3–8. https://doi.org/10.1080/10196789800000016

Tsalgatidou, A., & Pitoura, E. (2001). Business models and transactions in mobile electronic commerce: Requirements and properties. *Computer Networks,* 37(2), 221–236.

Turban, E., King, D., Lee, J. K., Liang, T.-P., & Turban, D. C. (2015). *Electronic commerce: A managerial and social networks perspective* (8th ed.). Springer.

Uckelmann, D., Harrison, M., & Michahelles, F. (2011). *Architecting the internet of things*. Springer. https://doi.org/10.1007/978-3-642-19157-2

Vaithianathan, S. (2010). A review of e-commerce literature on india and research agenda for the future. *Electronic Commerce Research*, 10(1), 83–97. https://doi.org/10.1007/s10660-010-9046-0

Vyas, D., Chisalita, C., & Dix, A. (2017). Organizational affordances: a structuration theory approach to affordances. *Interacting with Computers*, 29(2), 117–131.

Vygotsky, L. S. (1980). *Mind in society: Development of higher psychological processes*. Harvard University Press.

Wan, Y [Yan], Ma, B., & Pan, Y. (2018). Opinion evolution of online consumer reviews in the e-commerce environment. *Electronic Commerce Research*, 18(2), 291–311. https://doi.org/10.1007/s10660-017-9258-7

Wan, Y. (2015). The matthew effect in social commerce. *Electronic Markets*, 25(4), 313–324.

Wan, Y., Menon, S., & Ramaprasad, A. (2007). A classification of product comparison agents. *Communications of the ACM*, 50(8), 65–71.

Weber, R. (2012). Evaluating and developing theories in the information systems discipline. *Journal of the Association for Information Systems*, 13(1), 1–30.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Quarterly*, 26(2), xiii–xxiii.

Wirtz, B. W. (2018). *Electronic business* (6th ed.). Springer Gabler. http://www.springer.com/

Wortmann, F., & Flüchter, K. (2015). Internet of things. *Business & Information Systems Engineering*, 57(3), 221–224. https://doi.org/10.1007/s12599-015-0383-3

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

Xu, G., & Gutierrez, J. A. (2006). An exploratory study of killer applications and critical success factors in m-commerce. *Journal of Electronic Commerce in Organizations*, 4(3), 63–79. https://doi.org/10.4018/jeco.2006070104

Appendix

Appendix 2.1-A: Manifestations of IoT-commerce affordances along the customer buying process

Pre-purchase stage	
Step	**IoT-commerce affordance**
need recognition	▪ *Social interactions*: Recognizing an unidentified need through social interaction with peers ▪ *Personalized services*: Let oneself being guided by personalized product and service advertisement ▪ *Proactive services*: Let oneself being guided by proactive product and service recommendations ▪ *Location-based services*: Let oneself being guided by location-based advertisements and recommendations ▪ *Context-aware services*: Recognizing needs through environmental sensor data ▪ *Automated customer process*: Automated need recognition based on customer, peer, and public data
consideration	▪ *Electronic transactions*: Perceiving a lowered barrier to pursue the fulfillment of the need ▪ *Temporal independence*: Perceiving a lowered barrier to pursue the fulfillment of the need ▪ *Information transparency*: Perceiving a lowered barrier to pursue the fulfillment of the need ▪ *Social interactions*: Considering social knowledge and experiences of peers ▪ *Personalized services*: Let oneself being guided by personalized product and service advertisement ▪ *Proactive services*: Let oneself being guided by proactive recommendations ▪ *Spatial independence*: Perceiving a lowered barrier to pursue the fulfillment of the need ▪ *Location-based services*: Let oneself being guided by location-based advertisements and recommendations ▪ *Natural interactions*: Perceiving a lowered barrier to pursue the fulfillment of the need ▪ *Automated customer process*: Automated decision in the background whether to pursue the fulfillment of the need
search	▪ *Electronic transactions*: Searching a broad range of digital and non-digital products and services online ▪ *Temporal independence*: Searching products and services at any time ▪ *Online platforms*: Searching products and services across manufacturers on central marketplaces ▪ *Information transparency*: Getting access to a broad range of information about products, services, manufacturers, and other background information ▪ *Social interactions*: Considering social knowledge and experiences of peers ▪ *Personalized services*: Let oneself being guided by personalized search results ▪ *Proactive services*: Let oneself be guided by recommendations of certain products and services instead of search ▪ *Spatial independence*: Searching products and services from everywhere ▪ *Location-based services*: Let oneself being guided by search results based on location ▪ *Natural interactions*: Using voice, haptics, gesture, or other natural interaction to search for products and services ▪ *Automated customer process*: Replacing customer’s search by automated search in the background based on customer preferences
Purchase stage	
Step	**IoT-commerce affordance**
choice	▪ *Electronic transactions*: Choosing among relevant products and services online ▪ *Online platforms*: Accessing information about products and services via central marketplaces ▪ *Information transparency*: Data-based comparing of products and services ▪ *Social interactions*: Accessing user-generated content about products and services ▪ *Natural interactions*: Using voice, haptics, gesture, or other natural interaction to choose between products and services ▪ *Automated customer process*: Automated decision on product or service in the background based on customer preferences
ordering	▪ *Electronic transactions*: Ordering products and services remotely without the need to visit a brick and mortar store ▪ *Temporal independence*: Ordering products and services at any time ▪ *Online platforms*: Bundled ordering from a multitude of sellers via central marketplaces ▪ *Social interactions*: Sharing the moment of purchase and the whole shopping experience via social networks ▪ *Spatial independence*: Ordering products and services from everywhere ▪ *Natural interactions*: Using voice, haptics, gesture, or other natural interaction to order products and services ▪ *Automated customer process*: Replacing customer’s ordering by automated ordering in the background
payment	▪ *Electronic transactions*: Paying electronically with digital and non-digital currencies via the Internet ▪ *Temporal independence*: Paying orders at any time ▪ *Online platforms*: Bundled paying to a multitude of sellers via a marketplace ▪ *Social interactions*: Paying via P2P payment solutions ▪ *Natural interactions*: Using voice, haptics, gesture, or other natural interaction for payment ▪ *Automated customer process*: Replacing customer’s payment by automated payment in the background

Appendix 2.1-A: Manifestations of IoT-commerce affordances along the customer buying process (continued)

Post-purchase stage	
Step	**IoT-commerce affordance**
consumption/ usage	▪ *Electronic transactions*: Consuming/using digital content ▪ *Temporal independence*: Consuming/using digital content at any time ▪ *Online platforms*: Consuming/using digital content via a platform ▪ *Information transparency*: Getting access to data relevant for consumption/usage ▪ *Social interactions*: Sharing of consumption/usage experience with peers ▪ *Spatial independence*: Consuming/using digital content from everywhere ▪ *Location-based services*: Consuming/using digital content dependent on location ▪ *Context-aware services*: Accessing data of consumption/usage tracking ▪ *Natural interactions*: Using voice, haptics, gesture, or other natural interaction to steer consumption/usage
engagement	▪ *Electronic transactions*: Interacting via digital channels ▪ *Temporal independence:* Engaging at any time ▪ *Online platforms*: Sharing of user-generated content on the platform ▪ *Social interactions*: Engaging digitally with peers ▪ *Spatial independence* Engaging from everywhere ▪ *Context-aware services*: Sharing of context-aware data collected from IoT devices ▪ *Natural interactions*: Using voice, haptics, gesture, or other natural interaction for engagement ▪ *Automated customer process*: Replacing customers engagement via automated communication and data exchange with other connected devices, customers, and manufacturers
service requests	▪ *Electronic transactions*: Sending service requests online ▪ *Temporal independence*: Sending service requests at any time ▪ *Online platforms*: Having one unified point of contact for service requests ▪ *Social interactions*: Let oneself being guided by digital support from peers ▪ *Personalized services*: Receiving personalized customer service ▪ *Proactive services*: Let oneself being guided by proactive notices for service due ▪ *Spatial independence*: Sending service requests from everywhere ▪ *Location-based services*: Receiving location-based customer service ▪ *Context-aware services*: Initiating service requests through environmental sensor data ▪ *Natural interactions*: Using voice, haptics, gesture, or other natural interaction for service requests ▪ *Automated customer process*: Replacing customer’s service requests by automated service requests

Appendix 2.1-B: Textual explanation of manifestations of IoT-commerce affordances along the customer buying process

This Appendix B provides exemplifications for the manifestations of affordances along the buying process. The first step **need recognition** of the **pre-purchase stage** is explained in section 5. In the following, the remaining eight steps of the customer buying process are covered.

In the step '**consideration**', half of all affordances lead to lower perceived barriers to fulfill the recognized need due to the convenience of *electronic transactions* that customers may conduct *temporally and spatially independent.* Easy access to relevant information (*information transparency*) and *natural interactions* such as voice commands to voice assistants, haptic touches or clicks on a button of a replenishment service also contribute to a perceived lower barrier to fulfill one's need due to the knowledge that fulfillment with IoT devices is convenient and easy. Beyond the possibility that customers' consideration to pursuing the purchase of a product or service is guided by *personalized* advertisements and *proactive* recommendations that may even be *dependent on the location* of the IoT device and/or customer, *social interaction* may also reveal experiences of peers that influence the consideration whether and how to fulfill a need. Due to *customer process automation* with algorithms, IoT devices might also decide automatically whether to proceed in the buying process, without further intervention of the customer. For instance, smart energy management solutions already decide on their own whether a room shall be heated, or a lawn shall be watered or the washing machine decides automatically whether to reorder detergent.

The 'consideration' step is followed by the '**search**' step. Customers are able to search a broad range of products and services online (*electronic transactions*) and across manufacturers on online marketplaces (*online platforms*) *independent from temporal* and *spatial restrictions*. Search results can be tailored to the *location* and *personalized* to the preferences of an individual customer. Within this search, customers obtain easy access to information about products and services, its manufacturers, and other details relevant to the search (*information transparency*). Furthermore, social knowledge and experiences (*social interaction*), for instance, product ratings on a platform or social media posts from peers about the product, may guide the customer's search process. *Proactive* recommendations may influence and shorten the search process whereas smart algorithms of IoT devices may fully automate the search for customers (*automated customer process*). From an interaction point of view, voice assistants allow to conveniently search for products via voice commands, and replenishment services (e.g., smart fridges) via embedded touch displays (*natural interaction*).

The steps of the **purchase stage** (i.e., ‘**choice’**, ‘**ordering’**, and ‘**payment’**) comprise six to seven manifestations of affordances in each step. As known from e-commerce and m-commerce, choice, ordering, and payment in IoT-commerce can be conducted online via *electronic transactions*, *independent from temporal and spatial restrictions*, and via *online platforms*. *Information transparency* allows easy comparison of products and services that are enriched with *social interaction* such as the access to and the sharing of user-generated content in the steps ‘choice’ and ‘ordering’ or even the payment with a P2P payment solution such as Bitcoin. The main differences compared to the prior waves of e-commerce and m-commerce are driven by the IoT-commerce affordance *natural interaction* and *automated customer processes*. Customers can use *natural interaction* to choose between products and services, order, and pay them online. Voice assistants allow to execute commands via voice; smart resource management and replenishment services (e.g., smart fridges) offer displays for haptic touches; replenishment services such as the Amazon Dash Button allow to choose, order, and pay products online via a simplified interface – by pressing a button. In contrast, gesture control is not yet present in widespread IoT devices. Smart algorithms add intelligence to IoT devices and their connected service offerings (*automated customer process*). This intelligence enables IoT devices to choose, order, and pay products and services automatically based, for instance, on customer’s preferences. This automation is making any direct intervention of the customer obsolete, as for example fully self-sufficient replenishment services for washing detergent. Furthermore, solutions for smart resource management, however, already optimize the consumption of electricity, water, or other resources automatically without the customer’s involvement and therefore influence directly the quantity and time of resource consumption.

The steps of the **post-purchase stage** (i.e., ‘consumption/usage’, ‘interaction’, and ‘service request’) are located after the actual purchase stage and comprise eight to eleven affordances in each step. In the step of ‘**consumption/ usage**’, digital content (*electronic transactions*), for instance, a series on Netflix and music on Spotify, can be consumed via a platform (*online platforms*) at any time (*temporal independence*). Whenever you would like to listen to music on Spotify or watch a series on Netflix, you can log in to the platforms via any device (e.g., mobile phone or voice assistant) and instantly start to consume the product such as listening to a song. Due to the diversity of devices and the inherent nature of digital content, it can be consumed *spatially independent*. For instance, a song can be listened to via your Spotify account on your mobile phone during a picnic in the park or seamlessly in all rooms with connected voice assistants. Some providers offer *location-based services* during consumption/usage. For instance, a car-sharing vehicle displays car-sharing parking areas nearby or services that might initiate new buying processes such as vouchers for nearby shops. Furthermore, customers are able to share data, opinions, and experiences about consumption/usage of a product or service easily with peers, for instance via social networks (*social interactions*). It might be opinions about digital content, but also about non-digital products and services. Those written product reviews on platforms, shared moments of consumption/usage on Instagram and Facebook, and product ratings on the manufacturer's website become part of the user-generated content that might influence other peers in their buying process (e.g., within 'choice'). Furthermore, customers can get easy access to data relevant for consumption/usage of the product or service (*information transparency*). For instance, the name of the current song is displayed on the screen of the IoT device or the voice assistant is able to tell the ingredients of the bought food during cooking. IoT devices may also save historical data about consumption/usage, for instance, collected via sensors (*context-aware services*). Customers can get access to this background data about their consumption via a mobile app (e.g., energy consumption history provided by the smart resource management system). To steer the consumption of digital content, the customer can use *natural interactions* such as her voice. For instance, she can request the voice assistant to play a certain song, without having to search manually for it on the smartphone.

In the step of '**engagement**', customers can interact digitally (*electronic transactions*), at any time (*temporal independence*), and from everywhere (*spatial independence*) in order to engage with other customers (*social interaction*) or organizations, for instance, to provide feedback about products on platforms (*online platforms*), engaging in co-creation of a new product, or engaging in a digital customer community. Furthermore, customers can share and compare data collected by IoT devices via sensors with organizations and peer customers (*context-aware services*). Customers might want to compare their reduction of energy consumption since the purchase of a smart energy management system with historic consumption data of other customers. Active engagement of customers is facilitated via comfortable and easy interaction with connected devices in the living space of the customer. For instance, voice assistants (*natural interactions*) are integrated into customers' daily life – who can quickly leave a product rating of washing powder while doing the laundry. Smart algorithms replace individual customer actions via automated communication and data exchange among other connected devices, customers, and manufacturers (*automated customer processes*). For instance, smart thermostats already send data about energy consumption to the provider; nowadays mostly after allowed by the customer.

‘**Service requests**’ are sent online (*electronic transactions*) at any time (*temporal independence*) and from everywhere (*spatial independence*) with *natural interactions* as for instance voice. Due to online platforms, customers are offered one unified point of contact for sending service requests (*online platforms*): After logging into their account on a platform, customers see their previous service requests, can contact the provider and (re-)schedule appointments, without the necessity of using a second device, as for example a phone to call the service provider. In communities, customers receive support from their peers in problem solution (*social interaction*). Due to data about customer preferences, consumption history, sensor data, and other data available to the service provider, *personalized* and *location-based services* can be offered. Furthermore, customers receive notifications about service requests proactively (*proactive services*). Due to smart algorithms, service requests are even sent automatically without the involvement of the customer (*automated customer process*). An automated service request might be triggered through tracking data of the IoT device accessible to the provider, as for example a certain distance covered by the vehicle after which service is recommended, or sensor data (*context-aware services*), as for example a car automatically informing its sharing company in case of a breakdown.

2.2 IoT ethics – Status quo and directions for further research

Abstract:

As one of the most disruptive technology, IoT recently spreads in numerous areas of our lives and its potential is estimated with enormously high digits. Although discussions about ethics of emerging technologies are at the forefront, e.g., for artificial intelligence (AI), IoT specific ethical considerations remain crucial, but scarce. Based on a structured literature review, this research identifies and structures ethical issues of IoT, provides an overview of the current state of research for each aspect, and relates each aspect to IoTs features. Based on these results, concrete options for further research are provided, including suggestions for an examination of the transferability of known ethical issues from other technologies to IoT, identification of further ethical issues based on features and applications of IoT, a thorough examination of know ethical issues, and a positive view on IoT ethics. The results raise awareness for an intense examination of IoT ethics in research and practice.

Keywords: IoT, internet of things, ethical issues, structured literature review

Authors: Sarah Bayer

Status: This article is published in the Proceedings of the 42th International Conference on Information Systems (ICIS), December 12-15 2021, Austin, Texas.

2.2.1 Introduction

The potential of the Internet of Things (IoT) is described with a wealth of numbers: “McKinsey’s Global Institute predicts IoT will have an economic impact of between $4 trillion and $11 trillion by 2025” (McKinsey & Company, 2021). Other sources predict that “the Internet of things (IoT) will become a multi-trillion-dollar opportunity” (Shim et al., 2020) or that “it will grow to contain 64 to 73 billion connected devices by 2025” (Businessinsider, 2020; Shim et al., 2020). Those numbers are accompanied by enumerations of IoT’s benefits, e.g., efficiency, convenience, and personalization (IEEE, 2021). Examples of negative headlines about IoT are numerous: “The end of the autonomous consumer” (Die Zeit, 2016) “How your smart home devices can be turned against you” (BBC, 2020), “Unpatched Flaws in IoT Smart Deadbolt Open Homes to Danger” (threatpost, 2019), “Kids’ smart watches extremely vulnerable to being hacked” (babyology, 2019), or “Apple apologises for allowing workers to listen to Siri recordings” (The Guardian, 2019a). According to Avital et al. (2019), true impacts of IoT will only show over decades. Therefore, it is of utmost importance to be highly concerned with ethical issues that IoT can or could rise to avoid unintended consequences of IoT spread. Otherwise, the benefits of IoT are likely to be overwhelmed by its ethical issues.

Due to technology convergence, it is complex to address all layers of ethical issues of a technology. IoT is an excellent example for technology convergence, as it is usually defined as smart, hence embodying an artificial intelligence (AI) component. Although AI is frequently an inherent part of IoT, there is a lot in IoT that has nothing to do with AI. Elon Musk denotes AI as “humanity’s biggest existential threat” (Time, 2016), raising a host of ethical issues. There is a lot of awareness and research about ethical issues in AI (cf. B. C. Stahl et al. (2021) for an overview about ethical issues of AI) as well as calls for further research on AI ethics from leading players of research and media in this field (Ark, 2018; Kaplan & Haenlein, 2019; McKinsey, 2017), but a transfer of this research to the field of IoT is missing. Considering ethical issues for IoT is important as its potential is just beginning to unfold and will increasingly affect our lives in the years to come. It is crucial that this technology will not spread without severe ethical considerations. General applicability of existing research about ethical issues of AI or other emerging technologies to IoT is doubted (e.g., Cascone et al. (2017)). There is certainly an overlap of ethical issues between emerging technologies in general and IoT, but due to special features of IoT, as for instance sensors and actuators, the exact transferability and significance of ethical issues of other technologies to IoT have yet to be explored. Hence, in line with Bernd Carsten Stahl and Rogerson (2009) who define ICT ethics, I use the term IoT ethics "to denote ethical issues that arise from or in conjunction with" IoT.

In existing research, ethical problems of IoT mostly do not represent the core of the work, but are mentioned in passing without deeper examination (e.g., Avital et al., 2019; Bisaga et al., 2017). Solely Allhoff and Henschke (2018) provide an intense discussion of selected issues in IoT, stating that their discussions are "a first step […], fully aware that many more steps both should and will ensue" (Allhoff & Henschke, 2018, p. 56).

This study follows calls for further research about IoT ethics (e.g., Avital et al. (2019), Allhoff and Henschke (2018)). This paper argues that research in IoT ethics is crucial but scarce. It aims to identify and structure ethical issues of IoT that are discussed in literature. Furthermore, it provides a brief discussion of the current state of research for the respective issues and grounded on these results, proposes several directions for further research. Those objectives are accompanied by structuring IoT along its features (see theoretical background section) and illustrating ethical issues of IoT by application contexts. This paper is valuable as it raises awareness of the importance of research on IoT ethics. Additionally, it offers a structured overview and discussion of existing research in IoT ethics that does not exist yet, illustrating the importance of continuing research in IoT ethics, and offering concrete approaches for future research.

2.2.2 Theoretical background

Ethics, as part of philosophy, is a very large and ancient field of research. Hence, this paper cannot give a holistic and complete overview of it, neither of all current research streams of ethics related to digitalization. For this research aim, it is important to have a shared understanding of what ethics of technology and IoT ethics mean, and a broad overview of the main research streams of ethics and technologies.

Ethics can be defined as the philosophical study, or theoretical reflection of morality, or moral statements that search for the "grounds on which moral statements are made” (Bernd Carsten Stahl, 2012; Bernd Carsten Stahl & Rogerson, 2009; Vial, 2019). Ricoeur (1990) sees ethics as “the aim of a good life with and for each other” (Bernd Carsten Stahl et al., 2010, p.23). Moor (1985) states that ethics provides guidelines to determine what actions are good and what actions are bad. Often, ethics is “construed broadly, comprising not just what might be the philosophical dimension, but also the policy and legal components” (Allhoff & Henschke, 2018, p.56). Ethics in technology relates to “ensuring alignment with ethical norms” of technologies (European Comission, 2019, p.7). Due to its interdisciplinary nature, ethics of technology is examined from different perspectives: Technicians or IS-researchers often take a descriptive view (Moores & Chang, 2006), whereas philosophers frequently take a normative or metaethical perspective (Bynum, 2006; Luciano Floridi, 2005). From the philosophical perspective, the research question is part of information ethics, with its subcategory computer ethics emerging in the 1980s. This area of applied ethics examines ethical impacts of technologies (Bernd Carsten Stahl et al., 2010). No straight definition or categorization of this research stream exists (L. Floridi, 2004; Luciano Floridi, 2010). From the perspective of the IS community, ethics is part of the research stream about the dark sides of IS. Compared to the positive effects, as for instance rise of productivity and comfortability (Tian & S. X. Xu, 2015), this stream identifies, summarizes, and examines negative effects of the spread of technology in every part of human life. Mason (1986) was one of the first to examine ethical issues of IS, elaborating on privacy, accuracy, property, and accessibility. Most researchers pick out one or a few negative aspects and examine those in detail, e.g., privacy issues (Brown, 2000) or intellectual property issues (Burk, 2001). Few authors create an overview of existing issues (Kim et al., 2011; Pirkkalainen & Salo, 2016)), with Gimpel and Schmied (2019) summarizing existing overviews and creating a comprehensive taxonomy of risks and side effects of digitalization, ranging from personal- to socio-economic level. The term ethics is used in most of those publications, but some fuzzily consider all issues as ethical (Macnish et al., 2019), whereas others explicitly pick out some issues as ethical (Gimpel & Schmied, 2019). This circumstance reflects the unclear boundaries of ethics in interdisciplinary IS research. This paper does not aim to tease apart ethical and non-ethical issues. Looking at the definitions of ethics given above, it is clear that this alone would be a strongly philosophical research question. In this paper, what other authors have classified or called ethical is considered as an ethical issue.

This paper focuses on ethics of IoT, which belong to the category of emerging technologies. Especially for the development of emerging technologies with often unknown consequences, ethics play an important role in the use and spread (Bernd Carsten Stahl et al., 2010). Due to its nature, emerging technologies, as IoT, AI, or blockchain, are expected to further develop and expand during the next years. Therefore, it is unsure in which application areas the technology will expand, or how exactly one will use it. This makes it of utmost importance to push research about ethics already now, to avoid being overwhelmed by its speed of development, leaving ethical questions behind. So far, ethics of AI indisputably still occupies a significant part of the research stream ethics in technologies (Luciano Floridi, 2010). Next to philosophical perspectives on AI ethics (e.g., Rehg, 2014; Bernd Carsten Stahl, 2012), researchers examine one or few ethical issues, often in specific application contexts (e.g., Johnson et al., 2019; B. C. Stahl & Coeckelbergh, 2016). B. C. Stahl et al. (2021) provide a comprehensive categorization of ethics of AI, including metaphysical questions (e.g., superintelligence, change of human nature), general questions about living in a digital world (e.g., autonomy, distribution of benefits), and specific issues arising from machine learning (e.g., algorithmic biases, discrimination). Although AI is seen as one defining part of "smart connected products" (Porter & Heppelmann, 2014), those publications about ethics of AI and other emerging technologies do not sufficiently address IoT-specific features (Allhoff & Henschke, 2018).

Since the term IoT was used by the Massachusetts Institute of Technology (MIT) in 1999 (Cvijikj & Michahelles, 2011; Wortmann & Flüchter, 2015), there were set up numerous definitions of IoT (e.g. (Huber et al., 2017; McKinsey, 2017; Uckelmann et al., 2011). For the purpose of this paper, I aim for a definition that describes IoT's features, as I follow Bernd Carsten Stahl et al. (2010) and make the connection between features and potential ethical issues. Therefore, the following description of IoT, based on several publications, seems most appropriate for this paper:

IoT is a network that **connects uniquely identifiable** physical objects (that can be everyday objects such as refrigerators, etc.) to the internet (Avital et al., 2019; Bayer et al., 2021; IEEE, 2015; Kees et al., 2015; Porter & Heppelmann, 2014; Qadri et al., 2020; Shim et al., 2020; Weber, 2012) and is typically characterized by the following features:

- **connectivity**: interface between a source of data and a device (IEEE, 2015; Porter & Heppelmann, 2014)
- **ubiquity**: the network is available anywhere and anytime it is needed (IEEE, 2015)
- **sensing and actuating capabilities**: capability of identifying or recording features /device for moving or controlling that affects a physical entity (IEEE, 2015; Li et al., 2015; Porter & Heppelmann, 2014; Qadri et al., 2020; Shim et al., 2020)
- **intelligence**: embedded intelligence and knowledge functions (IEEE, 2015; Porter & Heppelmann, 2014)
- **communication capability**: protocols that enable communication (e.g. between device and cloud) (Allhoff & Henschke, 2018; Avital et al., 2019; Fletcher, 2016; IEEE, 2015; Wortmann & Flüchter, 2015)
- **programmability**: The device has a programmability feature, e,g., can initiate physical actions or processes (e.g. based on users commands or information from their

environment) and therefore enables automation or action-at-a-distance (Allhoff & Henschke, 2018; IEEE, 2015; Rosemann, 2013)

Due to its characteristics, IoT leads to a fusion of the digital and physical world (Huber et al., 2017). Application areas of IoT are diverse - structuring the field of applications is a research stream itself. For instance, based on Borgia (2014), Brandt et al. (2017) cluster the following application domains: individual well-being, smart city, smart energy, smart health, smart home, smart mobility. Equally based on Borgia (2014), Oberländer et al. (2018) name healthcare and public services on top of the previous named ones. Additionally, Wortmann and Flüchter (2015) name smart industry as a prominent application field of IoT. More detailed, Asghari et al. (2019) develop a comprehensive taxonomy of IoT applications comprising health care (e.g. smart wearables), environmental applications (e.g. smart agriculture), smart city (e.g. smart home, traffic monitoring), commercial applications (e.g., shopping systems), and industrial applications (e.g., smart grid). In summary, IoT is spreading into almost all areas of life, even though by no means all possible applications have already reached the final stage of development in everyday reality. Literature at the intersection of IoT and ethics is scarce. IoT literature often describes application areas (Asghari et al., 2019), business models (Dijkman et al., 2015), its architecture (Ray, 2018), or structures aspects of IoT in taxonomies (Oberländer et al., 2018). Regarding ethical issues in IoT, security and privacy are the most frequently named aspects (Suo et al., 2012). Publications are naming further IoT ethics (e.g., Avital et al., 2019), but not in a structured way and not explicitly connected to IoT applications or features. Additionally, those issues can often equally be attributed to digitalization or other emerging technologies, and no explanation is offered for the specific relevance of these issues for IoT. Solely Allhoff and Henschke (2018) offer a real discussion of single ethical aspects of IoT, but do not provide a holistic overview of IoT ethics. This paper aims to make a first step in the direction of filling this research gap via providing an overview of IoT ethics named in existing studies and offering a short discussion of each issue. Based on these insights, directions for further research in IoT ethics are derived.

2.2.3 Methods

To answer the research question, I reviewed in the first step existing literature about ethical issues of IoT. For this, I conducted a structured literature research with the following search term for titles, abstracts and keywords: *(“IoT”OR”Internet of things”)AND{(“ethic*”)OR(“moral*”)}*. Following the advice from Webster and Watson (2002), I included leading journals of IS (all Journals of the AIS Senior Scholars’ Basket of Eight[9] and the database AIS e-library). Furthermore, I searched within journals explicitly combining information or business with ethics (journals listed in the vhb jourqual 3[10], journals listed in FT50[11], journals listed in UT Dallas Research Ranking[12] explicitly dealing with ethics in economy, and journals identified from Bernd Carsten Stahl et al. (2010) dealing with information and ethics[13]). Due to the interdisciplinarity of the research question, I broadly expanded the search with the database ScienceDirect, covering “scientific, technical and medical research” (Elsevier B.V., 2021). The search resulted in 36 articles on which I conducted a title and abstract screening. An article was considered as relevant if it mainly dealt with IoT and if the abstract or title gave indications that examples, lists, or discussions of ethical issues or considerations are part of the paper. Therefore, papers that for instance named IoT purely as an example, but did not examine IoT in-depth, were considered as not relevant. After title and abstract screening, 20 articles were examined in-depth, meaning that the whole article was read and words, sentences, or paragraphs describing ethical issues of IoT were extracted. Note that as explained before, no personal evaluation of whether a marked aspect is actually ethical or not was made – this examination purely relied on the existing literature. In the end, 17 articles were identified as relevant for this research of ethics in IoT.

[9] Considered journals according to Association for Information Systems (2021): European Journal of Information Systems; Information Systems Journal; Information Systems Research; Journal of the AIS; Journal of Information Technology; Journal of MIS; Journal of Strategic Information Systems; MIS Quarterly.

[10] Considered journals according to VHB (2019): Journal of Business Ethics; Zeitschrift für Wirtschafts- und Unternehmensethik - Journal for Business, Economics & Ethics; Business Ethics: A European Review; Business Ethics Quarterly (BEQ).

[11] Considered journal according to Georgia State University (2021): Journal of Business Ethics.

[12] Considered journals according to The University of Texas at Dallas (2020): None.

[13] Considered journals according to Stahl et al. (2010): Ethics and Information Technology; Information, Communication and Society; International Review of Information Ethics; Journal of Information, Communication & Ethics in Society; Journal of Information Ethics; The Ethicamp Journal.

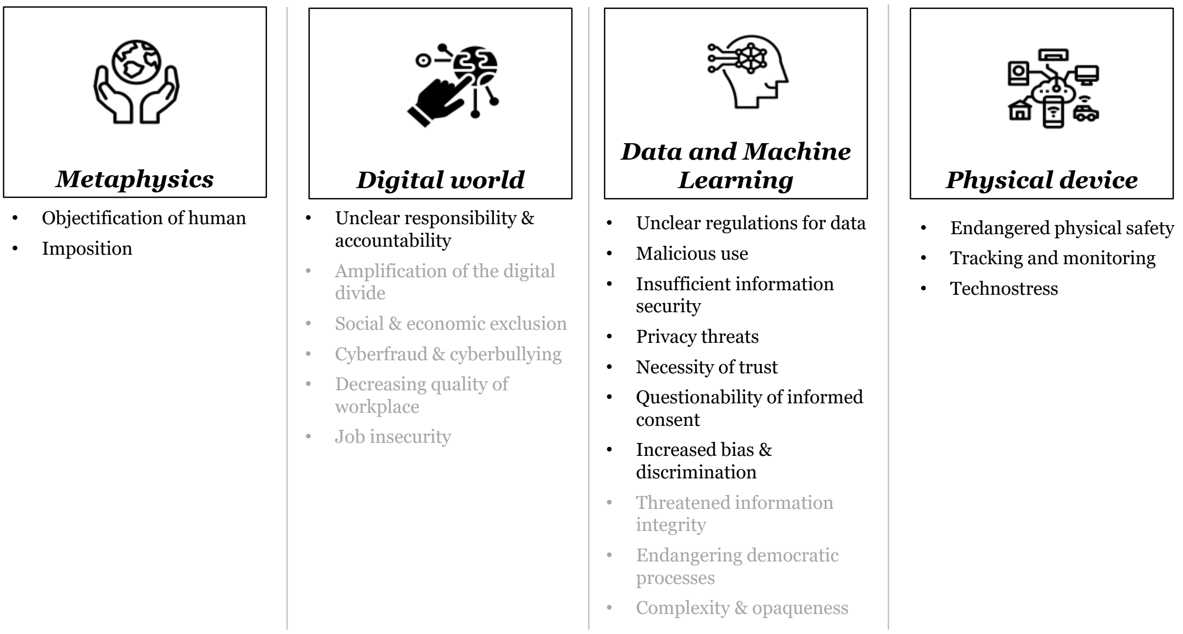
	# results	# relevant articles after title and abstract screening	# relevant articles after detailed examination
Journals: Basket of Eight	2	1	1
Database: AIS e-library	2	2	2
Journals: selection from vhb jourqual 3	0	0	0
Journals: selection from FT 50	0	0	0
Journals: selection from UT Dallas Research Ranking	0	0	0
Journals: advised from Stahl et al. (2010)	1	1	1
Database: ScienceDirect	19	16	13
#Results	36	20	17

Table 2.2-1: Overview of structured literature research

Afterward, the extracted words, sentences, or paragraphs were clustered and identical extracts were summarized to one issue. The resulting categories were inspired by the categorization of ethical issues of AI by B. C. Stahl et al. (2021). In the next chapter, I present the results of the structured literature research in Figure 2.2-1, followed by a short description, the current state of research as well as enumeration of the features of IoT primarily driving the respective issue, and exemplary applications illustrating the respective ethical issue.

2.2.4 Results

Four categories of IoT ethics are currently named or discussed in literature: metaphysical questions, general questions about life in a digital world, issues related to data and machine learning, and issues related to the physical device. The following figure presents each category with its respective ethical issues.



Metaphysics

- Objectification of human
- Imposition

Digital world

- Unclear responsibility & accountability
- Amplification of the digital divide
- Social & economic exclusion
- Cyberfraud & cyberbullying
- Decreasing quality of workplace
- Job insecurity

Data and Machine Learning

- Unclear regulations for data
- Malicious use
- Insufficient information security
- Privacy threats
- Necessity of trust
- Questionability of informed consent
- Increased bias & discrimination
- Threatened information integrity
- Endangering democratic processes
- Complexity & opaqueness

Physical device

- Endangered physical safety
- Tracking and monitoring
- Technostress

Figure 2.2-1: Ethical issues in IoT

Issues grayed out are named in existing literature as IoT ethics, but a particular relevance of this issues for IoT is not obvious. In the following, each issue specifically connected to IoT is described and an insight into the current state of research is provided. Although it is usually an interaction of several characteristics, IoT features that primarily push the respective issue are named. Note that features of IoT relate to the definition of IoT in the theoretical background section. Each issue is illustrated with application examples.

<u>Metaphysics</u>

Objectification of human (Calvo, 2020)

Description: Humans becoming connected things with their value measured by quantity of generated data.

State of Research: There is research about objectification stemming from the philosophical corner and predominantly in the health area (e.g. Cussins, 1996; Timmermans & Almeling, 2009), but there is no considerable research stream focusing on objectification of human due to IoT. Related, but on a higher level, discussions about AIs' evolution towards technological singularity and superintelligence with its detrimental effects on humanity take place (e.g., Matsumoto, 2018).

Features of IoT: Due to **connectivity** of **physical objects,** the technology itself integrates invisible into our lives, fostering objectification of human beings.

Applications: In a fully integrated smart home, the human itself is supported or replaced in numerous tasks from IoT, e.g., buying, showering, making coffee, controlling room temperature. Hence, user data is constantly collected during these tasks, enabling objectification.

Imposition (Calvo, 2020)

Description: Meaning is no longer provided by human needs, but by technical possibilities.

State of Research: Research about imposition of technology dates back several years and is often concerned with one technology or application scenario (e.g. e-readers (Thayer, A., Lee et al., 2011), higher education (Smith & Peck, 2020)). More recently, under the term "technological solutionism", Morozov (2015) critiques that technology often has solutions for problems that did not exist beforehand. Calvo (2020) names imposition as an issue in the context of IoT, but there is no current discussion or research stream.

Features of IoT: Especially **ubiquity** of IoT pushes unlimited innovation opportunities in numerous application contexts, which easily makes the question of meaningfulness recede into the background.

Applications: "Connected Dental Floss" uses data about user's daily tooth brushing habits to assign to the user the optimal length of dental floss (Schreier, 2018).

Digital world

Unclear responsibility & accountability (Avital et al., 2019; Calvo, 2020; Gill, 2016; Josephina & Andreas, 2019; U. Lee et al., 2019)

Description: Depersonalization and dissolution of responsibility and the associated accountability for actions or decisions taken by IoT devices.

State of Research: Connected to algorithmic decision-making embedded in technologies, there is research treating questions about accountability and justice for algorithmic decisions in general (e.g., Binns et al., 2018; Diakopoulos, 2016; Persson & Kavathatzopoulos, 2017), but without special focus on IoT features. Considering application areas, questions about responsibility are primarily examined in scenarios endangering physical safety, as for instance autonomous vehicles (Schuppli, 2014; Woldeamanuel & Nguyen, 2018) and healthcare (Grote & Berens, 2020).

Features of IoT: Due to embedded **intelligence** in IoT which might result in unexpected actions or incomprehensive decisions, as well as **programming** by software engineers and commands of users, the list of those possibly responsible is long (e.g., algorithm developers, software engineers, data scientists, users, IoT system) (Fritz et al., 2020).

Applications: In case of a crash with a self-driving car as it happened in 2018 with a test vehicle from Uber, there are many potential accounts, e.g., the car itself, the company with the employees who manufactured and programmed the car, or the safety-driver (Forbes, 2020).

Data and Machine Learning

Unclear regulations for data (Bisaga et al., 2017; Engin et al., 2020; Wen Shieng et al., 2018)

Description: No provision of a clear vision or regulation about IoT data ownership, collection, handling, and exploitation.

State of Research: Data handling is most often discussed within an architectural perspective on IoT (Ahad & Biswas, 2019; Singh et al., 2018; Srinivasan, 2018) and connected to privacy and security discussions (Amanullah et al., 2020; Jang et al., 2018; Shafagh et al., 2017; Varadharajan & Bansal, 2016, see explanations on security and privacy). Furthermore, there are authors specifically examining questions about data ownership for IoT without classifying it as an ethical problem (Janeček, 2018; Mashhadi et al., 2014).

Features of IoT: Due to its **sensing capabilities**, **connectivity,** and **communication capability**, a large amount of data can be collected and transferred.

Applications: Who owns data collected from carsharing vehicles – the driver, the car, the carsharing company?

Malicious use (Avital et al., 2019; Calvo, 2020; Josephina & Andreas, 2019; U. Lee et al., 2019)

Description: Unintended or intended unethical use and consequences of IoT system or IoT data. Closely connected to insufficient information security.

State of Research: Malicious use of technology in general is prominently in media due to well-known examples, e.g., Cambridge Analytics using data of 87 million Facebook users to set up advertising campaigns influencing the presidential elections in 2018 (The Guardian, 2018). In the context of IoT, Josephina and Andreas (2019) name encryption and anonymization as potential solutions to ensure a level of security avoiding malicious data usage. Furthermore, there is no considerable research stream about malicious use in IoT.

Features of IoT: Due to **ubiquity** of data collection and the **ability to communicate** those data, actions of malicious data usage can be very promising (but still unethical).

Applications: Embodying an implicit persuasion for product sales in smart home IoT devices that can support the buying process of the user is an example of malicious use of the IoT system through unethical use of user data (U. Lee et al., 2019).

Insufficient information security (Allhoff & Henschke, 2018; Biros, 2020; Cascone et al., 2017; Kopacek, 2018; Nikas et al., 2018; Ransbotham et al., 2016)

Description: Information security standards to avoid security incidents with IoT are ineffective and underdeveloped, which promotes related ethical issues, e.g., privacy, physical safety, trust.

State of Research: Together with privacy, security is the most frequently named issue of IoT. One reason might be that security promotes other, related ethical issues (e.g., privacy, trust, physical safety). There is a vast amount of research helping to design and implement secure technology, but due to neglecting hardware security and the potentially limited computing power of IoT, transferability to IoT is challenged (Cascone et al., 2017). Furthermore, Abdalla Ahmed, A.I., Ab Hamid, S.H. et al. (2019) state that existing security applications often fail for IoT, for instance as IoT devices can be captured physically and used as “a gateway to compromise the entire network” (Abdalla Ahmed, A.I., Ab Hamid, S.H. et al., 2019). Authors call for action towards more security of IoT, e.g., via better security education (Biros, 2020), introduce examples for security incidents (Nikas et al., 2018), or explain the cruciality of IoT-security (Cascone et al., 2017). More detailed, Allhoff and Henschke (2018) discuss the related user responsibility for security as well as the balance between usability and security. Next to academic research, media is equally pushing discussions about the lack of safety of IoT and its devastating consequences (The New York Times, 2016, 2019; Wired, 2016).

Features of IoT: **Sensors** can constantly collect data and the **ubiquity** of IoT devices makes it difficult for users to narrow down the space IoT devices occupy. Due to the **communication capability** of IoT devices, users recklessly lose control of their data.

Applications: IoT devices are used in distribute denial of service attacks to bring down websites (e.g., The Conversation Trust, 2016).

Privacy threats (Allhoff & Henschke, 2018; Avital et al., 2019; Biros, 2020; Calvo, 2020; Cascone et al., 2017; Engin et al., 2020; Josephina & Andreas, 2019; Kopacek, 2018; Nikas et al., 2018; Wachter, 2018; Weber, 2012; Wen Shieng et al., 2018)

Description: Endangered privacy due to collection and handling of a vast amount of data from IoT constantly surrounding individuals.

State of Research: Closely related to security, privacy is frequently mentioned as an IoT issue. Building on the vast number of publications that examine privacy in technologies, Allhoff and Henschke (2018) explain that privacy is an obvious problem for IoT since a very large amount and all kind of data is collected. Other researchers underline the importance of privacy in IoT (Weber, 2012), show how easily privacy can be threatened (Nikas et al., 2018), discuss the conflicting goals of privacy and identification for IoT (Wachter, 2018), or implement a privacy-respecting access control (Wen Shieng et al., 2018).

Features of IoT: With its **sensors** and **communication capability**, the IoT is the perfect medium to collect immense amount of sensitive data that has the potential to endanger individual's privacy on a new level. Due to **ubiquity**, technology such as smart home fades into the background and the user easily loses awareness of its presence and potential data collection.

Applications: Smart home devices collect data that can be analyzed for a change of relationship status, e.g., if a smart coffee machine is asked to make two instead of one coffee in the morning and the smart shower EvaDrop used more water than usual (Allhoff & Henschke, 2018). Employees of Apple listen to Siri’s records of user’s conversations to improve the technology (The Guardian, 2019b).

Necessity of trust (Allhoff & Henschke, 2018; Engin et al., 2020; Josephina & Andreas, 2019)

Description: Trust is particularly difficult to establish with IoT. Trust includes for instance the assumption that components of IoT "will reliably serve their function"(Allhoff & Henschke, 2018). Closely related to other ethical issues, as trust is only established if users have the feeling that other issues (e.g., security; physical safety) are sufficiently fulfilled.

State of Research: There is a vast amount of research about trust in IT (Gefen et al., 2008; Söllner et al., 2016) and specifically about trust in AI, often related to the research stream about explainable AI (Biran & Cotton, 2017; Miller, 2019), but no considerable research stream about trust in IoT.

Features of IoT: Trust is of special interest within IoT due to features as **ubiquity** and capability to **actuate** that require a particularly high level of user trust.

Applications: A fully integrated smart home that could potentially listen to every conversation within the appartement needs special trust from the user that privacy is guaranteed.

Questionability of informed consent (Allhoff & Henschke, 2018; Josephina & Andreas, 2019; Saarikko et al., 2020)

Description: It is questioned if consent within IoT is truly informed. Informed consent is defined as "any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her" (European Parliament and Council of the European Union, 2016).

State of Research: With its origin in medicine, informed consent was transferred to digital technologies in the early 2000s (Allhoff & Henschke, 2018; Eysenbach & Till, 2001). In the domain of digital technologies, it mostly refers to informed consent about collecting and handling data of the user. Frequently, the construct of informed consent is doubted due to intransparency and the high amount of explanations that most users do not understand or do not read before giving their consent. Especially for the vast amount of (highly sensitive) data collected through IoT, Allhoff and Henschke (2018) call for a truly informed consent for IoT, knowing that there must be a trade-off between the benefits and the costs of such an elaborated concept. Other researchers propose approaches for informed consent in IoT (Neisse et al., 2015; O'Connor et al., 2017).

Features of IoT: Due to **sensors** collecting data and their **capability of communication,** the user might not be aware of **ubiquitous** devices constantly collecting data, even though he or she once agreed to e.g., data collection when the device was put into operation for the first time.

Applications: Users once agree on terms and conditions when installing the smart speaker in their home, but are not aware of what their data is used for.

Increased bias & discrimination (Calvo, 2020; Engin et al., 2020; Josephina & Andreas, 2019; Ransbotham et al., 2016; Wachter, 2018)

Description: Systematization of biases and prejudices through algorithms embedded in IoT, leading to unjust treatment of different categories of people.

State of Research: Numerous publications are dealing with the risk "that technology enhances our human short- comings" (Persson & Kavathatzopoulos, 2017) and proposing solutions, e.g. laws and regulations or statistical analysis (Persson & Kavathatzopoulos, 2017). Often, publications examining bias and discrimination within IoT focus on the embedded AI (Tschider, 2018). Wachter (2018) states that profiling methods enabled by IoT foster discriminatory treatment, showing the importance of examining bias & discrimination specifically for IoT.

Features of IoT: Due to **sensors** constantly collecting a vast amount of data that can be **communicated** to data storage, big data analytics can reach another level of analysis that has the potential to be discriminating.

Applications: Data on users driving behavior collected in smart cars could lead to an increase in insurance premiums; data collected from smart cars could lead to structurally higher insurance premiums for women.

<u>Physical device</u>

Endangered physical safety (Allhoff & Henschke, 2018; Kopacek, 2018)

Description: IoT "has the potential to be active in physical realm" (Allhoff & Henschke, 2018) and therefore might endanger the physical safety of individuals with unexpected or unintended actions.

State of Research: There is awareness in research about cruciality of IoT-devices' physical safety (P. Lin et al., 2012; K. Lin et al., 2017), but research primarily focuses on the context of autonomous vehicles, in particular on discussions about how a vehicle should behave in situations where an accident is unavoidable (MIT Media Lab).

Features of IoT: Due to **actuators**, IoT devices can induce physical actions they are **programmed** for.

Applications: Autonomous car that runs over a child because it did not classify the child as a human being.

Tracking and monitoring (Biros, 2020; Wachter, 2018)

Description: IoT devices offer the potential to reveal habits and exceptions of individuals, facilitating monitoring from other parties.

State of Research: IoT research is mostly concerned with technical solutions, opportunities, and challenges of tracking and monitoring in healthcare (Hayati, N., & Suryanegara, M., 2017; Jung & Agulto, 2021; Patii & Iyer, 2017), road traffic (Anusha & Ahmed, 2017; Jisha et al., 2017) , or energy management (Kamienski et al., 2017).

Features of IoT: Due to **ubiquity** of IoT and its **sensing** capabilities, tracking and monitoring are effortless and extensive.

Applications: A smart lock at the front door reveals when the resident leaves the house; a smart coffee machine reveals when the user wakes up in the morning; a smart car reveals if the driver makes a detour or stopover on the way to work.

Technostress (Ransbotham et al., 2016)

Description: In this context, technostress refers to the stress experienced by individuals due to IoT use (Tarafdar et al., 2015). As IoT includes everyday objects such as refrigerators, lamps, or cars, two possible directions of technostress are imaginable: Either the technology fades into the background and the human being no longer perceives it as such, which leads to a reduction of technostress, or the feeling of constantly and always being surrounded by technology increases, resulting in higher technostress.

State of Research: There is a vast amount of research about technostress in IS literature examining technostress creators, consequences, and factors inhibiting technostress (e.g., Adam et al., 2017; Y.-K. Lee et al., 2014; Tarafdar et al., 2011), but no research specifically focus on IoT characteristics.

Features of IoT: Due to **ubiquity** of IoT's **physical objects**, users are always and everywhere surrounded by IoT.

Applications: Smart Home devices ubiquitously surround individuals in their homes, but as IoT is embedded in habitual physical objects, as for instance a fridge, it recedes into the background and might no longer be perceived primarily as technology.

Aspects that are named as ethical issues of IoT in literature but that do not have an obvious connection to features of IoT (and are therefore not explained in more detail) are the following:

- category “digital world”: amplification of the digital divide (Baiyere et al., 2020), social & economic exclusion (Calvo, 2020), cyberfraud & cyberbullying (Ransbotham et al., 2016), decreasing quality of the workplace (Cascone et al., 2017), job insecurity (Baiyere et al., 2020; Calvo, 2020; Kopacek, 2018)
- category “data and machine learning”: threatened information integrity (Baiyere et al., 2020), endangering democratic processes (Engin et al., 2020), complexity & opaqueness (Baiyere et al., 2020; Gill, 2016)

In sum, results of the literature research show that there are plenty of issues that are named as ethical issues in IoT. Apart from endangered physical safety, all of those issues can be equally found in discussions about ethical issues of emerging technologies in general, sometimes specifically focused on AI. Only for a fraction of these aspects, there is a real IoT-specific discussion, which considers features and applications of IoT.

2.2.5 Discussion

Obviously, the first three categories of IoT ethics (metaphysics, digital world, data and machine learning) do not originate solely from IoT-specific features, but from digitalization and spread of emerging technologies in general. A part of those aspects (aspects not greyed out in 2.2-1) play a special role in IoT, as IoT-specific characteristics strongly influence those issues, whereas others are not specifically connected to IoT (aspects greyed out in Figure 2.2-1). The fourth category includes issues emerging from the physical device component. Since the physical device is a major, increasing component of IoT, it is especially important to shed light on the issues associated with it.

Regarding features of IoT as introduced in the theoretical background section, ubiquity is most often the driving factor behind IoTs ethical issues. The IoT network being available anytime and anywhere is not yet always implemented, but the development of IoT goes in this direction. Currently, the limit of ubiquity is seen in managing the enormous amount of IoT devices, including data streams, leading to privacy and security questions (Vredenbregt, 2020). This is in line with the findings that ubiquity mainly reinforces issues of privacy and security, as well as the related issues of informed consent, tracking and monitoring, and trust. Usually, the feature of ubiquity is linked to communication capability. Without the option to transfer data collected via a device to other places, e.g., data storage, ubiquity itself solely stimulates technostress due to invasion of physical IoT devices constantly surrounding individuals and imposition due to unlimited innovation opportunities encouraged by the ubiquity of IoT.

Moreover, IoTs feature of sensing and actuating capability notably promotes ethical issues. The added value or defining characteristic of IoT devices is often seen precisely in sensors via which it can record sounds or temperatures, for example, and can execute actions autonomously via actuators (Asghari et al., 2019; Madakam et al., 2015; Porter & Heppelmann, 2014). Of course, other features as programmability and connectivity are needed for proper functioning of sensors and actuators, but users primarily perceive sensors and actuators. Sensors offer potential for privacy and security incidents, and elaborated tracking and monitoring. Furthermore, informed consent of data collected via sensors is questionable and data might induce biased and discriminatory consequences. Actuators primarily foster one, but major, point: physical safety. Compared to other digital technologies that are limited to their hardware, IoT has the potential to make physical changes in the environment (Allhoff & Henschke, 2018). Next to the perspective of endangered physical safety in IoT ethics, both in research and media there is a discussion about increased physical safety, e.g., via surveillance cameras, smart safes, reduction of car accidents (IoT for all, 2020; Kumar et al., 2021; Singhal et al.).

Concerning discussions of the identified ethical aspects, it is important to say that most authors solely enumerate the respective issue as applicable in their context of examination, but it does not comprise the main part of their paper. For instance, Bisaga et al. (2017) name data handling as an ethical issue in their context of smart energy, but without a thorough examination of this aspect. In line, Weber (2012) names privacy as an issue that should be considered in corporate social responsibility, but without a deeper discussion. Solely Allhoff and Henschke (2018) offer a true examination and discussion of informed consent, privacy, security, physical safety, and trust. For each aspect, they argue its cruciality for IoT and show its history in research. Apart from this publication, no other study has at its core the examination of IoT ethics. This shows the urgency of further, detailed, and rigor research in IoT ethics.

This paper contributes to research as it structures and categorizes literature of IoT ethics. Especially the category "physical device" raises awareness for IoT-specific research in ethics, as discussions about emerging technologies do not explicitly shed light on this category. Furthermore, this paper establishes a link between ethical issues and IoT's features, setting the base for further deep dive into IoT characteristics. In sum, our paper contributes via setting the hook for future research in IoT ethics, described in detail in the next section. For practice, our results show that development of IoT should include ethical aspects, especially concerning IoTs features of ubiquity, sensing- and actuating capability. Those features offer enormous potential for organizations, but in order to unfold, it is of utmost importance to address associated ethical issues. Privacy and security are well-known aspects, but due to the enormous amount of highly sensitive data that potentially can be collected through ubiquitous IoT devices, these aspects have to reach a new level. They should not be treated as inconvenient regulations, but as an decisive asset to successfully spread IoT. Companies should actively promote their privacy and security standards and features in order to gain trusting customers that extensively use IoT devices. Furthermore, physical safety is inherently connected to IoT's features that might intimidate customers. Through the possibility to make changes in the physical environment, users might be more reluctant to intensively use IoT compared to other technologies without this possibility. Hence, developers should clearly define and communicate the options and limits of physical intrusion of IoT devices. Without this, potential users might not build trust.

2.2.6 Limitations and further research

IoT is on the rise in virtually all areas of life, but as this research shows, the ethical issues raised specifically by the unique characteristics of this technology have yet to be explored in sufficient detail. This research structured existing discussions in literature to set the base for further research but left aside practical discussions about IoT ethics. As there probably are issues that are currently discussed in practice but not in scientific examinations, this constitutes a limitation of this research. Furthermore, the results include all issues that were named ethical from other authors, without questioning whether ethicist would truly attribute all of those aspects to ethics or not. Together with the results of this study, those limitations lead to the following areas for future research:

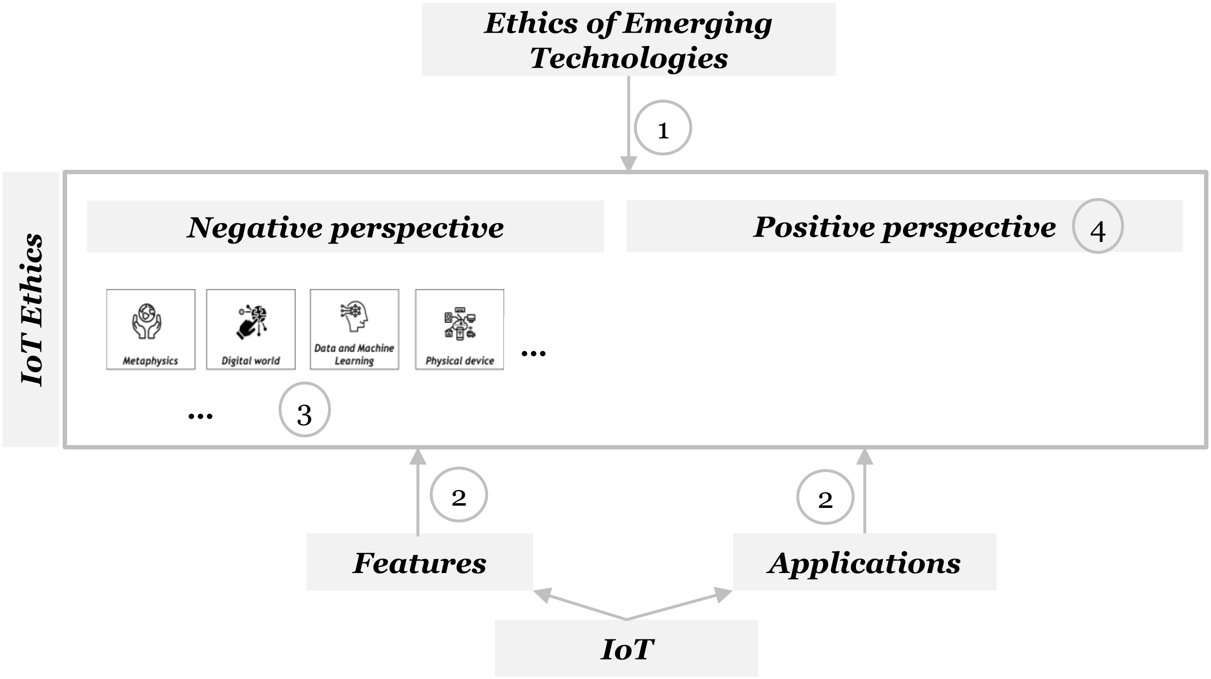


Figure 2.2-2: Future research areas for IoT ethics

(1) A thorough examination of transferability of known ethical issues from other technologies to IoT:

Existing research about ethical issues in technologies, as presented in the theoretical background section, should be examined for its **transferability to IoT**. Existing research about technology ethics can guide "similarly scholarly work in the IoT context" (Baiyere et al., 2020). As this research showed, most ethical issues of IoT can equally be found for other technologies, but there is a rare true examination of those issues specifically focusing on IoT features and application contexts. Therefore, further research can take those issues as a point of departure for studying their significance in IoT.

(2) Identification of further ethical issues based on features and applications of IoT:

As ethical issues are ubiquitous, a list of ethical issues of IoT can never be complete but is potentially infinite (Bernd Carsten Stahl et al., 2010). No one can be sure about all possible application contexts of IoT in the future and for that reason alone it is presumptuous to think that one could enumerate all ethical problems. This, of course, also applies to this study. Additionally, the structured literature research of this paper was comprehensive, but not all-encompassing. There might be ethical issues of IoT discussed in research that this search strategy was not able to cover. Nevertheless, this study provides an extensive overview of ethical issues of IoT currently named or discussed in literature. Further research could focus on identifying supplementary ethical issues of IoT, currently not yet discussed in literature. For this, researchers can follow advice from Bernd Carsten Stahl et al. (2010) and **analyze features of IoT or (futuristic) application areas of IoT** to expand Figure 2.2-1 with further aspects. Due to diversity of IoT devices, it is a challenge to examine IoT ethics without focusing on a specific application context, as most researchers currently do. Nevertheless, as the examination showed, analyzing features of IoT for ethical issues brings a holistic perspective and interesting, structured results. Thus, further research can take features as a point of departure to examine IoT ethics and therefore provide a useful basis for an overarching, academic discussion detached from individual application examples. As IoT constantly evolves, it is important to identify potential ethical issues as early as possible to avoid pure reactive behavior on issues, and foster proactivity.

(3) A thorough examination of known ethical issues of IoT:

There needs to be an awareness that IoT requires a separate research stream. This stream should include the **in-depth analysis** of known issues transferred to IoT (e.g., security, privacy, responsibility, and accountability) and the emergence of issues that are seen as primarily raised in IoT (e.g., physical safety). Furthermore, as shown before, **ubiquity, communication capability, and sensing and actuating** capabilities are the features of IoT that mainly cause ethical problems. Hence, a special focus on further research should lie on those features. Detailed examination of their impact on existing and emerging IoT devices is crucial to take sufficient account of the uniqueness of IoT compared to other technologies.

(4) A positive view of IoT on ethics:

It might be interesting to examine whether IoT can make a positive contribution to ethics by attenuating known issues from other technologies. This research for instance raised the question about IoT's influence on technostress that might be positive or negative (see result section). This impact could be further examined, e.g., in an experimental setting with fully integrated smart home devices.

2.2.7 Conclusion

This work was motivated by analyzing discussions about ethical issues in the context of IoT. With an extensive structured literature research, the relevant literature was identified, analyzed and the results were summarized in four categories with the respective ethical issues of IoT named in existing literature. For each issue, a description of the issue and the current state of research was provided. IoT features primarily responsible for each issue were assigned and application examples, illustrating the issue, were given. The results show that research in IoT ethics is indispensable to ensure a responsible expansion of IoT in all areas of human life. Current research in IoT ethics is too scarce and not focused on IoT-specific features. IoT is likely to become one of the most important technologies in near future. Profound research in this area is needed to support and guide a meaningful proliferation of IoT, whose ethical vulnerabilities will not eventually overtake us.

References

Abdalla Ahmed, A.I., Ab Hamid, S.H., Gani, A., Suleman K., and Khan, M. K. 2019. Trust and reputation for Internet of Things: Fundamentals, taxonomy, and open research challenges. *Journal of Network and Computer Applications* (145), p. 102409.

Adam, M. T. P., Gimpel, H., Maedche, A., and Riedl, R. 2017. Design Blueprint for Stress-Sensitive Adaptive Enterprise Systems. *Business & Information Systems Engineering*, 59(4), 277-291.

Ahad, M. A., and Biswas, R. 2019. Request-based, Secured and Energy-efficient (RBSEE) Architecture for Handling IoT Big Data. *Journal of Information Science*, 45(2), 227-238.

Allhoff, F., and Henschke, A. 2018. The Internet of Things: Foundational ethical issues. *Internet of Things*, 1-2, 55-66.

Amanullah, M. A., Habeeb, R. A. A., Nasaruddin, F. H., Gani, A., Ahmed, E., Nainar, A. S. M., and Imran, M. 2020. Deep learning and big data technologies for IoT security. *Computer Communications*, 151, 495-517.

Anusha, A., and Ahmed, S. M. 2017. Vehicle Tracking and Monitoring System to Enhance the Safety and Security Driving Using IoT. *Proceedings of the International Conference on Recent Trends in Electrical, Electronics and Computing Technologies*, 49-53.

Ark, T. V. 2018. *Let’s talk about AI ethics. We are on a deadline*. https://www.forbes.com/sites/tomvanderark/2018/09/13/ethics-on-a-deadline/# 220a07602e21. Accessed 8 March 2021.

Asghari, P., Rahmani, A. M., and Javadi, H. H. S. 2019. Internet of Things applications: A systematic review. *Computer Networks*, 148, 241-261.

Association for Information Systems 2021. *Senior Scholars' Basket of Journals*. https://aisnet.org/page/SeniorScholarBasket. Accessed 1 May 2021.

Avital, M., Dennis, A. R., Rossi, M., Sørensen, C., and French, A. 2019. The Transformative Effect of the Internet of Things on Business and Society. *Communications of the Association for Information Systems*, 44(1), 129-140.

babyology 2019. *Kids’ smart watches extremely vulnerable to being hacked, experts warn*. https://babyology.com.au/health/safety/kids-smart-watches-extremely-vulnerable-to-being-hacked-experts-warn/. Accessed 3 February 2021.

Baiyere, A., Topi H., Venkatesh, V., Wyatt, J., and Donnellan, B. 2020. The Internet of Things (IoT): A Research Agenda for Information Systems. *Communications of the Association for Information Systems,* 47(21).

Bayer, S., Gimpel, H., and Rau, D. 2020. IoT-commerce - opportunities for customers through an affordance lens. *Electronic Markets*, 31(1), 27–50.

BBC 2020. *How your smart home devices can be turned against you*. https://www.bbc.com/future/article/20200511-how-smart-home-devices-are-being-used-for-domestic-abuse. Accessed 3 February 2021.

Binns, R., van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. 2018. It's Reducing a Human Being to a Percentage. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1-14.

Biran, O., and Cotton, C. 2017. Explanation and justification in machine learning: A survey. in *XAI workshop at the 26th International Joint Conference on Artificial Intelligence*, Melbourne, Australia, 8-13.

Biros, D. 2020. The Challenges of New Information Technology on Security, Privacy and Ethics. *Journal of the Midwest Association for Information Systems Journal of the Midwest Association for Information Systems,* 2(1).

Bisaga, I., Holford, N., Grealish, A., Baker-Brian, C., and Parikh, P. 2017. Scalable Off-Grid Energy Services Enabled by IoT: a Case Study of BBOXX SMART Solar. *Energy Policy* (109), 199-207.

Borgia, E. 2014. The Internet of Things vision: Key features, applications and open issues. *Computer Communications* (54), 1-31.

Brandt, R., Püschel, L., Röglinger, M., and Schlott, H. 2017. Unravelling the internet of things: A multi-layer taxonomy and archetypes of smart things. *Working Paper of the FIM Research Center.*

Brown, W. S. 2000. Ontological Security, Existential Anxiety and Workplace Privacy. *Journal of Business Ethics*, 23(1), 161-165.

Burk, D. L. 2001. Copyrightable functions and patentable speech. *Communications of the ACM*, 44(2), 69-75.

Businessinsider 2020. *The Internet of Things 2020: Here's what over 400 IoT decision-makers say about the future of enterprise connectivity and how IoT companies can use it to grow* *revenue.* https://www.businessinsider.com/internet-of-things-report?IR=T. Accessed 8 March 2021.

Bynum, T. W. 2006. Flourishing Ethics. *Ethics and Information Technology*, 8(4), 157-173.

Calvo, P. 2020. The ethics of Smart City (EoSC): moral implications of hyperconnectivity, algorithmization and the datafication of urban digital society. *Ethics and Information Technology*, 22(2), 141-149.

Cascone, Y., Ferrara, M., Giovannini, L., and Serale, G. 2017. Ethical issues of monitoring sensor networks for energy efficiency in smart buildings: a case study. *Energy Procedia*, 134, 337-345.

Cussins, C. 1996. Ontological Choreography: Agency through Objectification in Infertility Clinics. *Social studies of science*, 26(3), 575-610.

Cvijikj, I., and Michahelles, F. (eds.) 2011. *The Toolkit Approach for End-User Participation in the Internet of Things*. Berlin, Heidelberg: Springer.

Diakopoulos, N. 2016. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2), 56-62.

Die Zeit 2016. *Internet der Dinge: Das Ende des souveränen Konsumenten*. https://www.zeit.de/2016/12/kuenstliche-intelligenz-internet-der-dinge-konsumenten-kaufentscheidung/seite-2?utm_referrer=https%3A%2F%2Fwww.google.de%2F. Accessed 7 February 2021.

Dijkman, R. M., Sprenkels, B., Peeters, T., and Janssen, A. 2015. Business models for the Internet of Things. *International Journal of Information Management*, 35(6), 672-678.

Elsevier B.V. 2021. *Science Direct*. https://www.sciencedirect.com/. Accessed 14 February 2021.

Engin, Z., van Dijk, J., Lan, T., Longley, P. A., Treleaven, P., Batty, M., and Penn, A. 2020. Data-driven urban management: Mapping the landscape. *Journal of Urban Management*, 9(2), 140-150.

European Comission 2019. *Ethics Guideline for Trustworthy AI*. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Accessed 12 February 2021.

European Parliament and Council of the European Union 2016. Parliament and Council of the European Union, Regulation (EU) 2016/679 (General Data Protection Regulation)

Eysenbach, G., and Till, J. E. 2001. Ethical issues in qualitative research on internet communities. *British Medical Journal* (323), 1103-1105.

Fletcher, D. 2016. *Internet of Things*. in Evolution of Cyber Technologies and Operations to 2035, M. Blowers (ed.), Cham: Springer International Publishing, pp. 19-32.

Floridi, L. 2004. *The Blackwell guide to the philosophy of computing and information*, Malden, MA: Blackwell Pub.

Floridi, L. 2005. Information ethics, its nature and scope. *Computer and Society*, 35(3).

Floridi, L. 2010. *The Cambridge Handbook of: Information and Computer Ethics*, Cambridge, UK: Cambridge University Press.

Forbes 2020. *Who Is Responsible In A Crash With A Self-Driving Car?* https://www.forbes.com/sites/fernandezelizabeth/2020/02/06/who-is-responsible-in-a-crash-with-a-self-driving-car/?sh=7a44b00e4b2b. Accessed 9 March 2021.

Fritz, A., Brandt, W., Gimpel, H., and Bayer, S. 2020. Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI). *De Ethica: A Journal of Philosophical, Theological and Applied Ethics*, 6(1), 3-22.

Gefen, D., Benbasat, I., and Pavlou, P. 2008. A Research Agenda for Trust in Online Environments. *Journal of Management Information Systems*, 24(4), 275-286.

Georgia State University 2021. *Executive Doctoral in Business: Financial Times's top Journals*. https://research.library.gsu.edu/c.php?g=388883&p=2638798. Accessed 1 February 2021.

Gill, K. S. 2016. Data Driven Wave of Certainty- a question of ethical sustainability. *IFAC-PapersOnLine*, 49(29), 117-122.

Gimpel, H., and Schmied, F. 2019. Risks and Side Effects of Digitalization: A Multi-Level Taxonomy of the adverse Effects of Using Digital Technologies and Media. *Proceedings of the 27th European Conference on Information Systems (ECIS)*, Stockholm & Uppsala, Sweden.

Grote, T., and Berens, P. 2020. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, 46(3), 205-211.

Hayati, N., & Suryanegara, M. 2017. The IoT LoRa system design for tracking and monitoring patient with mental disorder. *IEEE International Conference on Communication, Networks and Satellite*, 135-139.

Huber, R., Püschel, L., and Röglinger, M. 2017. IoT-enabled Smart Service Systems - Identification of Actors and Interaction Types. *Working Paper of the FIM Research Center*.

IEEE 2015. *Towards a Definition of Internet of Things (IoT)*. https://iot.ieee.org/images/files/pdf/IEEE_IoT_Towards_Definition_Internet_of_Things_Revision1_27MAY15.pdf. Accessed 5 March 2021

IEEE 2021. *6 Ways You'll Directly Benefit from the Internet of Things*. https://innovationatwork.ieee.org/6-iot-benefits/. Accessed 5 March 2021.

IoT for all 2020. *IoT Security and Physical Safety*. https://www.iotforall.com/iot-security-and-physical-safety. Accessed 5 March 2021.

Janeček, V. 2018. Ownership of personal data in the Internet of Things. *Computer Law & Security Review*, 34(5), 1039-1052.

Jang, J., Jung, I. Y., and Park, J. H. 2018. An effective handling of secure data stream in IoT. *Applied Soft Computing* (68), 811-820.

Jisha, R. C., Jyothindranath, A., and Kumary, L. S. 2017. IoT based school bus tracking and arrival time prediction. *International Conference on Advances in Computing, Communications and Informatics*, 509-514.

Johnson, K., Pasquale, F., and Chapman, J. 2019. Artificial Intelligence, Machine Learning, and Bias in Finance: Toward Responsible Innovation. *Fordham Law Review*, 88(2), 499.

Josephina, A., and Andreas, A. 2019. Case Study The Internet of Things and Ethics. *The ORBIT Journal*, 2(2), 1-29.

Jung, Y., and Agulto, R. 2021. A Public Platform for Virtual IoT-Based Monitoring and Tracking of COVID-19. *Electronics*, 10(1).

Kamienski, C. A., Borelli, F. F., Biondi, G. O., Pinheiro, I., Zyrianoff, I. D., and Jentsch, M. 2017. Context design and tracking for IoT-based energy management in smart cities. *IEEE Internet of Things Journal*, 5(2), 687-695.

Kaplan, A., and Haenlein, M. 2019. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1), 15-25.

Kees, A., Oberländer, A. M., Röglinger, M., and Rosemann, M. 2015. Understanding the Internet of Things: A Conceptualisation of Business-to-Thing (B2T) Interactions. *Proceedings of the 23rd European Conference on Information Systems (ECIS).*

Kim, W., Jeong, O. R., Kim, C., and So, J. 2011. The dark side of the Internet: Attacks, costs and responses. *Information Systems*, 36(3), 675-705.

Kopacek, P. 2018. Development Trends in Cost Oriented Production Automation. *IFAC-PapersOnLine*, 51(30), 39-43

Kumar, V., Ramachandran, D., and Kumar, B. 2021. Influence of new-age technologies on marketing: A research agenda. *Journal of Business Research* (125), 864-877.

Lee, U., Han, K., Cho, H., Chung, K.-M., Hong, H., Lee, S.-J., Noh, Y., Park, S., and Carroll, J. M. 2019. Intelligent positive computing with mobile, wearable, and IoT devices: Literature review and research directions. *Ad Hoc Networks* (83), 8-24.

Lee, Y.-K., Chang, C.-T., Lin, Y., and Cheng, Z.-H. 2014. The Dark Side of Smartphone Usage: Psychological Traits, Compulsive Behavior and Technostress. *Computers in Human Behavior* (31), 373-383.

Li, S., Da Xu, L., and Zhao, S. 2015. The internet of things: a survey. *Information Systems Frontiers*, 17(2), 243-259.

Lin, K., Abney, K., and Jenkins, R. 2017. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, New York, NY: Oxford University Press.

Lin, P., Abney, K., and Bekey, G. 2012. *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press Cambridge.

Macnish, K., Ryan, M., and Stahl, B. 2019. Understanding Ethics and Human Rights in Smart Information Systems. *The ORBIT Journal*, 2(2), 1-34.

Madakam, S., Ramaswamy, R., and Tripathi, S. 2015. Internet of Things (IoT): A Literature Review. *Journal of Computer and Communications*, 3(5), 164-173.

Mashhadi, A., Kawsar, F., and Acer, U. G. 2014. Human data interaction in IoT: The ownership aspect. *IEEE world forum on Internet of Things*, 159-162.

Mason, R. O. 1986. Four Ethical Issues of the Information Age. *MIS Quaterly*, 10(1), 5-12.

Matsumoto, T. 2018. *The Day AI Becomes God: The Singularity Will Save Humanity*, Cambridge, New Zealand: MEDIA TECTONICS.

McKinsey 2017. *Artificial intelligence: the next digital frontier?* Discussion Paper.

McKinsey & Company 2021. *Internet of Things*. https://www.mckinsey.com/featured-insights/internet-of-things/how-we-help-clients. Accessed 8 March 2021.

Miller, T. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* (267), 1-38.

MIT Media Lab. *Moral Machine*. https://www.moralmachine.net/. Accessed 5 January 2021.

Moor, J. 1985. What is computer ethics? *Metaphilosophy*, 16(4), 266-275.

Moores, T. T., and Chang, J. C.-J. 2006. Ethical Decision Making in Software Piracy: Initial Development and Test of a Four-Component Model. *MIS Quaterly*, 30(1).

Morozov, E. 2015. To save everything, click here. The folly of technological solutionism, New York: Public Affairs.

Neisse, R., Baldini, G., Steri, G., Miyake, Y., Kiyomoto, S., and Biswas, A. R. 2015. An agent-based framework for Informed Consent in the internet of things. *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, 789-794.

Nikas, A., Alepis, E., and Patsakis, C. 2018. I know what you streamed last night: On the security and privacy of streaming. *Digital Investigation* (25), 78-89.

O'Connor, Y., Rowan, W., Lynch, L., and Heavin, C. 2017. Privacy by Design: Informed Consent and Internet of Things for Smart Health. *Procedia Computer Science* (113), 653-658.

Oberländer, A. M., Röglinger, M., Rosemann, M., and Kees, A. 2018. Conceptualizing Business-to-Thing Interactions - A Sociomaterial Perspective on the Internet of Things. *European Journal of Information Systems*, 27(4), 486-502.

Patii, N., and Iyer, B. 2017. Health monitoring and tracking system for soldiers using Internet of Things (IoT). *International Conference on Computing, Communication and Automation*, 1347-1352.

Persson, A., and Kavathatzopoulos, I. 2017. How to make decisions with algorithms: Ethical decision-making using algorithms within predictive analytics. *ACM Computers & Society*, 47(4).

Pirkkalainen, H., and Salo, M. 2016. Two Decades of the Dark Side in the Information Systems Basket: Suggesting Five Areas for Future Research. *ECIS 2016 Proceedings*.

Porter, M. E., and Heppelmann, J. E. 2014. How smart, connected products are transforming companies. *Harvard Business Review*.

Qadri, Y. A., Nauman, A., Zikria, Y. B., Vasilakos, A. V., and Kim, S. W. 2020. The Future of Healthcare Internet of Things: A Survey of Emerging Technologies. *IEEE Communications Surveys & Tutorials*, 22(2), 1121-1167.

Ransbotham, S., Fichman, R. G., Gopal, R., and Gupta, A. 2016. Special Section Introduction—Ubiquitous IT and Digital Vulnerabilities. *Information Systems Research*, 27(4), 834-847.

Ray, P. P. 2018. A survey on Internet of Things architectures. *Journal of King Saud University - Computer and Information Sciences*, 30(3), 291-319.

Rehg, W. 2014. Discourse ethics for computer ethics: A heuristic for engaged dialogical reflection. *Ethics and Information Technology*, 17(1), 27-39.

Ricoeur, P. 1990. *Soi-même comme un autre*, Paris: Seuil.

Rosemann, M. 2013. The internet of things: New digital capital in the hands of customers. *Business Transformation Journal*.

Saarikko, T., Westergren, U. H., and Blomquist, T. 2020. Digital transformation: Five recommendations for the digitally conscious firm. *Business Horizons*, 63(6), 825-839.

Schreier, J. 2018. *Die lustigsten (sinnlosesten?) IoT-Devices der Welt*. https://www.industry-of-things.de/die-lustigsten-sinnlosesten-iot-devices-der-welt-a-689525/. Accessed 10 March 2021.

Schuppli, S. 2014. Deadly Algorithms: Can Legal Codes Hold Software Accountable for Code that Kills? *Radical Philosophy* (178), 2-8.

Shafagh, H., Burkhalter, L., Hithnawi, A., and Duquennoy, S. 2017. Towards blockchain-based auditable storage and sharing of IoT data. *Proceedings of the 2017 on Cloud Computing Security Workshop*, 45-50.

Shim, J. P., Sharda, R., French, A. M., Syler, R. A., and Patten, K. P. 2020. The Internet of Things: Multi-faceted Research Perspectives. *Communications of the Association for Information Systems*, 46, 511-536.

Singh, A., Garg, S., Batra, S., Kumar, N., and Rodrigues, J. J. 2018. Bloom filter based optimization scheme for massive data handling in IoT environment. *Future Generation Computer Systems*, 82, 440-449.

Singhal, A., Sarishma, and Tomar, R. Intelligent accident management system using IoT and cloud computing. 2016. *2nd International Conference on Next Generation Computing Technologies*, 89-92.

Smith, A., and Peck, B. 2020. The teacher as the ‘digital perpetrator’: Implementing web 2.0 technology activity as assessment practice for higher education Innovation or Imposition? *Procedia-Social and Behavioral Sciences*, 2(2), 4800-4804.

Söllner, M., Benbasat, I., Gefen, D., Leimeister, J. M., and Pavlou, P. 2016. Trust. *MIS Quarterly Research Curations*.

Srinivasan, A. 2018. IoT cloud based real time automobile monitoring system. *3rd IEEE International Conference on Intelligent Transportation Engineering*, 231-235.

Stahl, B. C. 2012. Morality, Ethics, and Reflection: A Categorization of Normative IS Research. *Journal of the Association for Information Systems*, 13(8).

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Laulhé Shaelou, S., Patel, A., Ryan, M., and Wright, D. 2021. Artificial intelligence for human flourishing – Beyond principles for machine learning. *Journal of Business Research*, 124, pp. 374-388.

Stahl, B. C., and Coeckelbergh, M. 2016. Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems*, 86, 152-161.

Stahl, B. C., Heersmink, R., Goujon, P., Flick, C., van den Hoven, J., Wakunuma, K., Ikonen, V., and Rader, M. 2010. Identifying the Ethics of Emerging Information and Communication Technologies. *International Journal of Technoethics*, 1(4), 20-38.

Stahl, B. C., and Rogerson, S. 2009. Landscapes of Emerging ICT Applications in Europe. *Proceedings of the Eighth International Conference of Computer Ethics: Philosophical Enquiry*.

Suo, H., Wan, J., Zou, C., and Liu, J. 2012 - 2012. Security in the Internet of Things: A Review. in 2012 I*nternational Conference on Computer Science and Electronics Engineering*, Hangzhou, Zhejiang, China. 23.03.2012 - 25.03.2012, IEEE, 648-651.

Tarafdar, M., D'Arcy, J., Turel, O., and Gupta, A. 2015. The dark side of information technology. *MIT Sloan Management Review*, 56(2), 61.

Tarafdar, M., Tu, Q., Ragu-Nathan, T. S., and Ragu-Nathan, B. S. 2011. Crossing to the Dark Side: Examining Creators, Outcomes, and Inhibitors of Technostress. *Communications of the ACM*, 54(9), 113.

Thayer, A., Lee, C. P., Hwang, L. H., Sales, H., Sen, P., and Dalal, N. 2011. The imposition and superimposition of digital reading technology: the academic potential of e-readers.

Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2917-2926.

The Conversation Trust 2016. *Hacked by your fridge: the Internet of Things could spark a new wave of cyber attacks*. https://theconversation.com/hacked-by-your-fridge-the-internet-of-things-could-spark-a-new-wave-of-cyber-attacks-66493. Accessed 5 January 2021.

The Guardian 2018. *Cambridge Analytica: how did it turn clicks into votes?* https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie. Accessed 5 January 2021.

The Guardian 2019a. *Apple apologises for allowing workers to listen to Siri recordings.* https://www.theguardian.com/technology/2019/aug/29/apple-apologises-listen-siri-recordings. Accessed 5 January 2021.

The Guardian 2019b. *Apple contractors 'regularly hear confidential details' on Siri recordings.* https://www.theguardian.com/technology/2019/jul/26/apple-contractors-regularly-hear-confidential-details-on-siri-recordings. Accessed 5 January 2021.

The New York Times 2016. *Stepping Up Security for an Internet-of-Things World.* https://www.nytimes.com/2016/10/17/technology/security-internet.html. Accessed 5 January 2021.

The New York Times 2019. *In the Rush to Join the Smart Home Crowd, Buyers Should Beware.* https://www.nytimes.com/2019/01/22/business/smart-home-buyers-security-risks.html. Accessed 12 March 2021.

The University of Texas at Dallas 2020. *The UTD Top 100 Business School Research Rankings.* https://jindal.utdallas.edu/the-utd-top-100-business-school-research-rankings/list-of-journals. Accessed 5 January 2021.

threatpost 2019. *Unpatched Flaws in IoT Smart Deadbolt Open Homes to Danger*. https://threatpost.com/unpatched-flaws-in-iot-smart-deadbolt-open-homes-to-danger/146871/. Accessed 5 January 2021.

Tian, F., and S. X. Xu 2015. How do Enterprise Resource Planning Systems Affect Firm Risk? Post-Implementation Impact. *MIS Quaterly*, 39(1), 39-60.

Time 2016. *What 7 of the World's Smartest People Think About Artificial Intelligence*. https://time.com/4278790/smart-people-ai/. Accessed 8 March 2021.

Timmermans, S., and Almeling, R. 2009. Objectification, standardization, and commodification in health care: A conceptual readjustment. *Social Science & Medicine*, 69(1), 21-27.

Tschider, C. A. 2018. *Regulating the internet of things: discrimination, privacy, and cybersecurity in the artificial intelligence age*. Denver Law Review, 96(87).

Uckelmann, D., Harrison, M., and Michahelles, F. 2011. *Architecting the internet of things*, Berlin, Heidelberg: Springer.

Varadharajan, V., and Bansal, S. 2016. Data security and privacy in the internet of things (iot) environment. *Connectivity Frameworks for Smart Devices*, 261-281.

VHB 2019. *VHB-JOURQUAL 3*. https://vhbonline.org/vhb4you/vhb-jourqual/vhb-jourqual-3. Accessed 14 February 2021.

Vial, G. 2019. Understanding digital transformation: A review and a research agenda. The *Journal of Strategic Information Systems*, 28(2), 118-144.

Vredenbregt, A. 2020. *Challenges put by the ubiquity of IoT devices*. https://bits-chips.nl/artikel/challenges-put-by-the-ubiquity-of-iot-devices/. Accessed 5 January 2021.

Wachter, S. 2018. Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR. *Computer Law & Security Review*, 34(3), pp. 436-449.

Weber, R. H. 2012. Corporate social responsibility as new challenge for the IT industry. *Computer Law & Security Review*, 28(6), 634-640.

Webster, J., and Watson, R. T. 2002. Guest Editorial: Analyzing the Past to Prepare for the Future: Writing a literature Review. *MIS Quarterly*, 26(2), xiii-xxiii.

Wen Shieng, P. S., Jansen, J., and Pemberton, S. 2018. Fine-grained Access Control Framework for Igor, a Unified Access Solution to The Internet of Things. *Procedia Computer Science*, 134, 385-392.

Wired 2016. *The internet of things will turn our machines against us*. https://www.wired.co.uk/article/internet-of-hackable-things. Accessed 5 February 2021.

Woldeamanuel, M., and Nguyen, D. 2018. Perceived benefits and concerns of autonomous vehicles: An exploratory study of millennials’ sentiments of an emerging market. *Research in Transportation Economics*, 71, 44-53.

Wortmann, F., and Flüchter, K. 2015. Internet of Things. *Business & Information Systems Engineering*, 57(3), 221-224.

3 Behind the scenes of Artificial Intelligence

3.1 Fear of algorithms: A synopsis of concerns about automated decision-making

Abstract:

Automated decision-making (ADM) is making its impact in all areas of modern life. Decisions previously made by humans are increasingly supported or replaced by algorithms. Many people harbor reservations about ADM, and yet, there is no exhaustive study that structures these concerns. The objective of our research is to outline a comprehensive framework of concerns about ADM. Based on a structured review of the literature and a qualitative content analysis of semi-structured interviews, we identified ten major concerns regarding the underlying technology, data, or the decision itself. Furthermore, we identified 14 concerns about the potential consequences of using ADM. Our framework is intended to guide future research on concerns about ADM, while also serving as a touchstone for anyone developing ADM-related offers and services that account for the potential reservations of the intended user group.

Keywords: algorithm, automated decision-making, algorithmic decision-making, concerns

Authors: Sarah Bayer, Fabian Schmied, Daniela Waldmann

Status: This article is a working paper.

3.1.1 Introduction

Algorithms are “a sequence of computational steps that transform inputs into outputs, and range from simple if-then statements to artificial intelligence (AI), machine learning, and neural networks” (Martin, 2019). Nowadays, algorithms are involved in all areas of life, for instance by producing news articles based on structured data, by supporting recruitment processes, by detecting fraud in sports betting, by deciding which physicians see which patient, and by defining dynamic prices in many application areas, such as e-commerce (e.g., Amazon), tourism (e.g., Airbnb), and transportation (e.g., Uber) (Diakopoulos, 2016; Martin, 2019; van den Broek et al., 2019). In some of these areas, we see “complex and networked algorithms that are beyond proper human understanding and control” (Gimpel & Schmied, 2019, p. 8). This comes with certain adverse, unexpected, and unintended effects (Gimpel & Schmied, 2019; Majchrzak et al., 2016), and these effects – positive as well as negative – are extending their reach into all aspects of modern life (Diakopoulos, 2016). Decision-making processes previously made by humans are increasingly supported (augmented by technology) or even replaced by algorithms (fully automated) (Martin, 2019; Wachter et al., 2017).

In the future, algorithms are expected to gain even more influence due to an ever-increasing degree of automation in decision-making processes as well as the expansion of application areas of ADM. This is affecting individuals, organizations, and society at large. On the one hand, organizations and public authorities may benefit from the accuracy, scale, speed, simplicity, and cost-efficiency of automated decisions (Diakopoulos, 2016). There are those who argue that algorithmic decisions foster objectivity and fairness (van den Broek et al., 2019). Others predict that algorithms may have significant negative consequences for individuals affected by automated decisions. Two prime examples are when potentially biased algorithms support policing (known as predictive policing) or assist judicial decision-making in court (Angwin et al., 2016; Binns et al., 2018; Corbett-Davies et al., 2016; Dressel & Farid, 2018; Martin, 2019). Algorithmic decisions are further criticized for facilitating other ethical violations such as sexism or privacy invasions (van den Broek et al., 2019). In this paper, we focus on reservations that individuals harbor about ADM.

Prior literature has already investigated potential risks and side effects of ADM for individuals, such as discrimination, lack of data protection, unfairness, or wider ethical issues. Most research articles discussed these issues in highly specific (and primarily future-oriented) use cases. However, there is no comprehensive overview of the chief concerns held by individuals when dealing with ADM, which is a necessary foundation to improve ADM adoption. What is missing, therefore, is a synopsis of these concerns about ADM derived from literature (focused mainly on specific single use cases) and complemented with a survey of multiple ADM cases. To fill this gap in the research and to provide a starting point for further, detailed research about these concerns, we aim to answer the following question:

Which concerns do individuals have about the use of automated decision-making?

The overview we have generated in reply to this question may serve as a foundation upon which others can develop responsible and transparent ADM-related offers and services with full regard for the fears and reservations of those affected (Diakopoulos, 2016). Furthermore, we intend to summarize as well as extend existing research to offer a basis for future research.

The paper is structured as follows: The following section provides the theoretical background for algorithmic decision-making and concerns. Then, we describe the methodological approach of our structured literature search and the qualitative content analysis of our semi-structured interviews, followed by the presentation of results. After the result section, or discussion includes practical and theoretical implications, and an outlook towards future research, followed from the conclusion.

3.1.2 Theoretical background

To understand concerns about ADM one must first dive deeper into the negative aspects of IT. Although there is an apparent pro-IT bias in information systems (IS) research, there is also research on the "dark side of IT." The Information Systems Journal published two consecutive special issues on the dark side of information technology use (Tarafdar et al., 2015a, 2015b). These special issues comprise articles that focus on one negative aspect of IT use at a time, such as technostress, IT interruptions, computer abuse, IT-mediated control, or unauthorized file sharing (Tarafdar et al., 2015a, 2015b). Further, Pirkkalainen and Salo (2016) review 37 articles in the AIS Senior Scholars' Basket of Journals and detect four types of dark side phenomena: information overload, IT addiction, and IT anxiety. Kim et al. (2011) provide a taxonomy of the dark side of the internet and focus on attacks, costs, and appropriate responses. They identify technology-centric dark side effects like spam, malware, hacking, and digital property rights violations. Additionally, they identify non-technology-centric dark side effects such as online theft, cyberbullying, and the aiding and abetting of crime. Gimpel and Schmied (2019) aim to provide a broad overview of dark side phenomena by developing a taxonomy of the most severe risks and side effects of digitalization, such as adverse exchange, adverse economic shifts, impairment of health, undesirable behavioral adaptation, or losing control over algorithms. Some of those dark sides of IT also relate to the use of algorithms.

ADM takes place when a result, e.g., a recommendation or a purchase, is achieved without human intervention (Allen & Masters, 2019). Thus, ADM is either supported by modern information and communication technologies (ICTs) or the decision is entirely made by the application of specific algorithms (Allen & Masters, 2019). This is why ADM is also called algorithmic decision-making. Another way to achieve ADM, however, may be to use complex artificial intelligence (AI) supported and trained by machine learning (ML) (Allen & Masters, 2019). Within AI, the different analytical techniques, such as descriptive, predictive, or prescriptive analytics, facilitate ever greater intelligence and business efficiency. Whereas descriptive and predictive analytics require a human manager to interpret the results, prescriptive analytics enables ADM (Vahn, 2014). In other words, it goes beyond predicting future results by anticipating what will happen, when it will happen, and why it will happen. What is more, it gives recommendations that benefit from those predictions (Kumar, 2015; Shankararaman & Gottipati, 2015). Consequently, prescriptive analytics answers the question "How can we make it happen?" (Shankararaman & Gottipati, 2015).

The impact of ADM on the lives of individuals triggers certain concerns about ADM. According to Lowry et al. (2011), we define concerns in use cases of ADM as the extent to which a person worries about possible risks and consequences associated with ADM use. The existing literature has already discussed the concerns some individuals have about ADM, e.g., discrimination (Strobel, 2019), or data privacy (Newell & Marabelli, 2015). It has also discussed factors that inhibit ADM adoption, e.g., control (Dietvorst et al., 2018) or trust (Castelo et al., 2019). It has further discussed a variety of use cases for ADM, e.g., automated travel planning (Cho & Han, 2019), autonomous driving (Dietrich & Weisswange, 2019), or automated purchases (Ringe et al., 2019), and the literature has also already discussed the implementation of ADM in business use cases (Dwivedi et al., 2021). However, these discussions have typically been grouped around single concerns, and most of the studies have focussed on a specific context. A comprehensive overview of concerns that might inhibit ADM adoption does not yet exist. In this paper, we argue for the need of a better understanding of how individuals perceive the impact of using ADM in daily life. This is necessary if we are to gain a deeper insight into the relationship between the perception of ADM's use, the perception of the consequences of ADM's use, and actual behavior, because organizations need to know which consequences individuals fear and how to address those negative perceptions (Karwatzki et al., 2017).

Further areas of academic research, such as data privacy, has indicated that individual concerns can be manifold (Hauff et al., 2015; Smith et al., 1996). Smith et al. (1996) have identified seven major data privacy concerns of customers (including data collection, secondary use, or improper access). Hauff et al. (2015) have investigated how perceived privacy-invasive data collection and usage can affect individuals. Their research has shown that, for some individuals, there are concerns at different levels. Meanwhile, Karwatzki et al. (2017) have developed a categorization of how individuals perceive the consequences of access to their personal information. This categorization spans seven types of consequences: psychological, social, career-related, physical, resource-related, prosecution-related, and freedom-related. Nevertheless, this research has merely discussed data privacy concerns (e.g., regarding unauthorized access to individuals' information), which we believe to be only one type of concern about ADM. As such, the existing research does not provide a comprehensive overview of potential concerns.

3.1.3 Research methodology and approach

To answer our research question, we take a two-step approach by way of a structured literature search and a qualitative content analysis of semi-structured interviews. First, we reviewed the existing (IS) literature to identify concerns about ADM. In so doing, we also identified current use cases for ADM, which served as a basis for the semi-structured interviews conducted in the second step. We used a search string, combining "automated decision" with the most common synonym used in the literature ("algorithmic decision"), as well as the term "prescriptive analytics," which is used primarily in the research area of statistics. Furthermore, we linked those expressions with "concern" and synonyms for concern commonly used in the literature, which yielded the following search terms: *("automated decision" OR "algorithmic decision" OR "prescriptive analytics") AND ("concern" OR "risk" OR "attitude" OR "danger" OR "aversion")*. As advised by Webster and Watson (2002), we did not restrict our literature search to databases with a focus on the IS discipline (covered by the databases ACM Digital Library and AIS Electronic Library). Instead, we expanded our search to general databases so as to cover a wide range of different research areas with our main focus directed at the domain of electronic commerce and computer science, engineering, law, marketing, logistics, and beyond (covered by the databases Science direct, EBSCOhost, JSTOR Library, SpringerLink, ProQuest). Since ADM is frequently embedded in highly topical discussions about AI, we included news from associations and academic journals. The structured literature search resulted in 175 articles. After the initial screening of titles and abstracts, the full texts of the remaining 30 articles were examined, whereupon 18 articles were classified as relevant. An article was considered relevant if the following two conditions were met: (1) the article dealt with ADM in general or in a specific use case and (2) the article named or explained concerns or adverse effects of ADM for a specific use case or in general terms. With regard to those 18 articles, we highlighted words or phrases expressing concerns about ADM (e.g., “discrimination” (Strobel, 2019), “computer implementation may be incorrect” (Brauneis & Goodman, 2018)) and use cases for ADM (e.g., “recommender systems” (Borràs et al., 2014), “loan application” (Strobel, 2019)).

This also proved to be highly useful in preparing the semi-structured interviews, which we then conducted to identify further concerns about ADM. We chose interviewees with diverse backgrounds to cover a broad cross-section of the population in terms of age and gender as well as educational and professional backgrounds. We met the interviewees in person or spoke to them on video calls, and in each case we recorded the interview. In total, we conducted 13 interviews, as shown in Table 3.1-1.

ID	Age	Gender	Highest educational level	Profession / Occupation
1	25	male	University degree	Student
2	60	female	High school diploma	Secretary
3	28	male	University degree	Doctoral candidate
4	34	male	Secondary school	IT administrator
5	33	female	University degree	Doctoral candidate
6	29	male	University degree	Technical employee
7	26	female	Secondary school	Nurse
8	28	male	University degree	Student
9	27	male	University degree	Doctoral candidate
10	57	male	Secondary school	Civil servant
11	57	female	University degree	Civil servant
12	22	male	High school diploma	Student
13	28	female	University degree	Doctoral candidate

Table 3.1-1: Demographic overview of interviewees

After 11 interviews, the 12th did not reveal further insights of any relevance. We conducted a 13th interview anyway, but this, too, revealed nothing new. Reassured that we had reached saturation point, we determined that we had gathered enough data via interviews. The duration of each ranged from 15 to 45 minutes and comprised four steps: (1) present a definition of ADM (“decisions that are made or at least supported by algorithms”) and ensure a common understanding of ADM, (2) ask open questions about prior experiences with ADM and any associated concerns, (3) present five use cases to discuss concerns with regard to each use case, (4) present and discuss the results of our literature search.

We presented five ADM use cases (automated lending (Brauneis & Goodman, 2018), intelligent travel bots (Cho & Han, 2019), automated evaluation of applicants (Faliagka et al., 2012), autonomous driving (Dietrich & Weisswange, 2019), and automated purchases (Ringe et al., 2019)). We chose those use cases because they are current in both mainstream media and academic research, and because they cover a broad spectrum of modern life, ranging from consumption, travel and locomotion, to the professional environment. Furthermore, we attached importance to the fact that the cases represent current progress as well as future scenarios. We provided the interviewees with images and a short description of these use cases. We transcribed all interviews verbatim in order to conduct a qualitative content analysis in line with the eight steps proposed by Schreier (2013). These eight steps bring together the best of various approaches to a thorough qualitative content analysis (Boyatzis, 1988; Hsie & Shannon, 2005; Mayring, 2010; Rustemeyer, 1992). We used the software MAXQDA to code the interviews, and each step of this methodology is outlined in detail below.

(1) Deciding on a research question: Our research question was defined ahead of the interviews (cf. Section 1).

(2) Selecting material: We conducted semi-structured interviews, each of which was fully transcribed. As our interview sample includes two different types of stakeholders (students and doctoral candidates involved in ADM research as well as individuals without professional experience in ADM), we chose two interviews from each group in order to set up the coding frame.

(3) Building a coding frame: To build main categories (“structuring”) and generate the subcategories (“generating”), we combined a concept- and data-driven approach. Since our ultimate aim is to analyze concerns about ADM, the main category of the coding frame is *concerns about ADM*. In the following, where we only use one main category, we also refer to categories on the second level as main categories, while categories on the third level are called sub-categories. The results of our literature research were used to generate certain main- and sub-categories in a concept-driven way (e.g., technology, data and societal as main categories, as opposed to privacy incidents, discrimination and job loss as sub-categories). Furthermore, we adopted the strategy of subsumption as proposed by Mayring (2010) for data-driven categories: We reviewed the interview transcripts until we encountered a relevant aspect, then checked whether this aspect is already covered by a category and either attributed the aspect to the existing category or created a new category (e.g., organizational for main categories, as opposed to lack of enjoyment and lack of spontaneity for subcategories).

As advised by Schreier (2013), our coding frame meets the requirements of unidimensionality (our main categories are unidimensional), mutual exclusiveness (sub-categories within one main category are mutually exclusive), and exhaustiveness (all relevant aspects of the material are covered by a category). After the definition of the coding frame, we defined each category (Schreier, 2013). Subsequently, we examined the bigger picture of the coding frame, then merged and split a few categories, and refined our definitions.

(4) Segmentation: As suggested by Schreier (2013), we divided our material into segments. Since the use cases of ADM mentioned in the interviews are suitable to specify the start and the end of a unit, we chose the use cases as a thematic criterion for segmentation.

(5) Trial coding: In the next step, we applied the coding frame to further interview transcripts. We split the material among the researchers and each researcher coded the material twice within two weeks.

(6) Evaluating and modifying the coding frame: We evaluated consistency and validity. Less than 10% of codes were assigned to different categories in two coding rounds. We discussed the respective categories and revised each definition. As we did not have any leftover categories but managed to assign each code to a proper category, we determined our coding frame to be valid. See Table 3.1-2 and 3.1-3 for the coding frame.

(7) Main analysis: We coded the rest of the interviews, and due to the high validity and consistency, there was no need to double-code the rest of the material (Schreier, 2013).

(8) Presenting and interpreting the findings: Below, we present our framework in visual terms alongside explanations of the categories of concerns in Tables 3 and 4. Additionally, we explain each category, illustrated by quotes in the following section.

3.1.4 Results

With the help of our structured literature search and semi-structured interviews, we identified 24 concerns. 13 concerns resulted from the structured literature search, 22 from the semi-structured interviews, which is to say that eleven emerged from both sources. Figure 2.2-1 structures the 24 concerns. We divided the framework into two categories of concerns: On the left-hand side of the chart, we identify concerns inherent to technology, data, or decisions. Those concerns do not necessarily have a direct impact but can develop into graver concerns about the consequences on the right-hand side.

Since applied technology, such as an algorithm, needs data to make automated decisions for the user, the concerns on the left-hand side of the framework are divided into three categories:

technology, data, and decision. These concerns about technology, data, and decision can lead to further concerns in different categories adapted from Karwatzki et al. (2017) and described in Table 3.1-2.

Category	**Definition**
Physical	Loss of physical safety due to the application of ADM
Social	Change in social status due to the application of ADM
Resource-related	Loss of resources due to the application of ADM
Psychological	Negative impact on one's peace of mind due to the application of ADM
Prosecution-related	Legal actions taken against an individual due to the application of ADM
Career-related	Negative impacts on one's career due to the application of ADM
Freedom-related	Loss of freedom of opinion and behavior due to the application of ADM

Table 3.1-2: Categories of concerns about consequences that individuals have due to the use of ADM adapted from Karwatzki et al. (2017)

A concern on the left-hand side can give rise to more than one concern on the right-hand side. For example, *"poor decision quality"* can lead to various specific concerns at different levels on the right-hand side of the framework, e.g., *"negative financial impact"* if the algorithm opts for more expensive consumer goods, *"negative physical impact"* if the autonomous driving car gets involved in an accident, or *"discrimination"* if the algorithm discriminates females for job offers. With the icons in Figure 3.1-1, we indicate whether a concern originates from semi-structured interviews (microphone) and/or from the literature review (book).

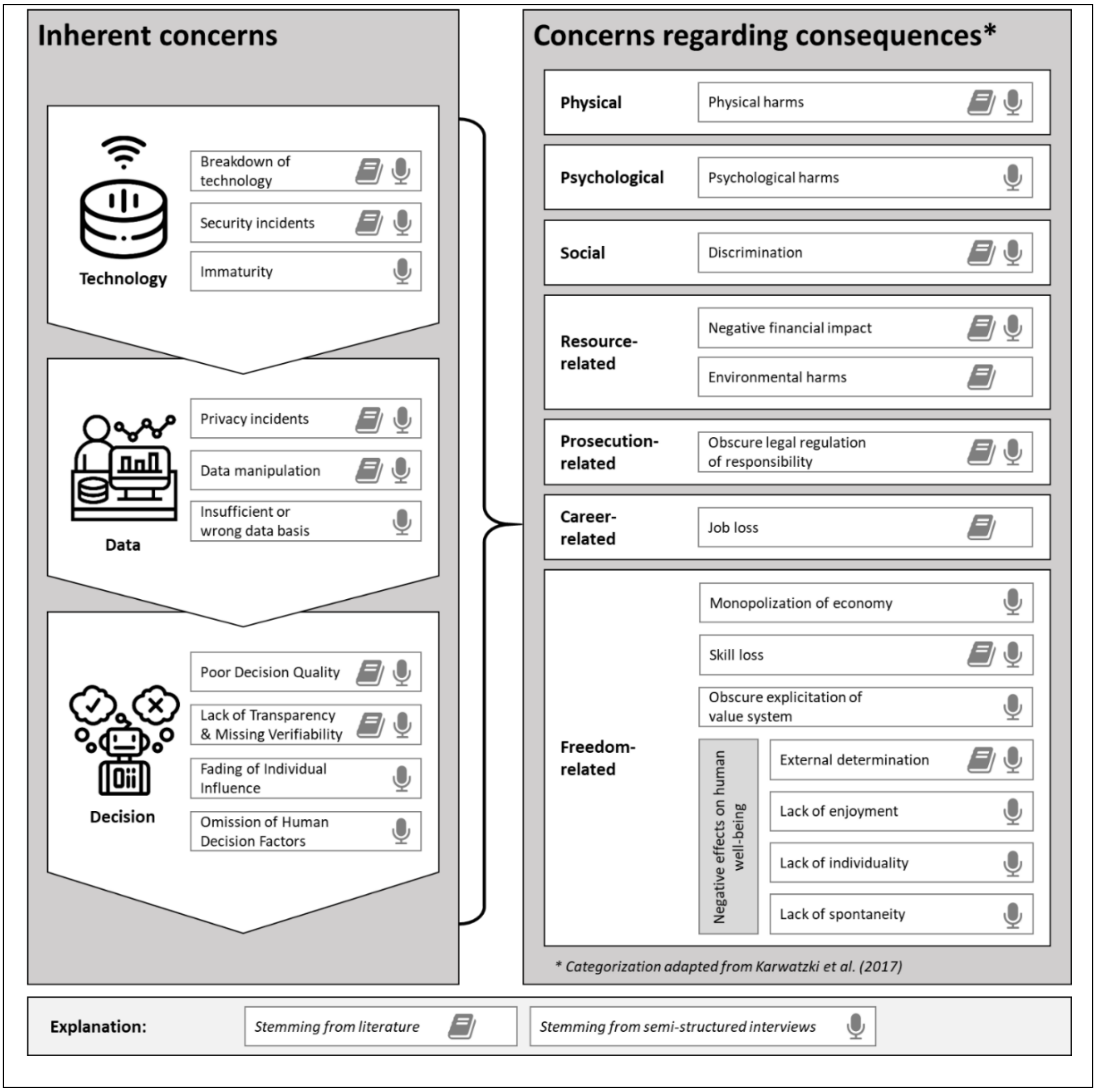


Figure 3.1-1: Framework of concerns about the use of ADM

Job loss and *environmental harms* are the only two aspects that did not occur in any interview but solely in the literature. All other 13 concerns that we found in the literature were confirmed in the interviews. Furthermore, our interviews added four concerns to the framework's left-hand side and seven concerns to the right-hand side. Table 3.1-3 presents the inherent concerns (left-hand side of Figure 3.1-1). Table 3.1-4 presents the concerns about consequences (right-hand side of Figure 3.1-1). For each concern, an explanation is provided, and literature sources as well as the IDs of the respective interviewees are shown to identify the origin of each concern.

Concerns	**Description**	**Literature sources**	**Interviews**
Technology			
Breakdown of technology	*Concerns about failures in technology or single features of technology*	Winters, 2017; Woldeamanuel & Nguyen, 2018	5, 10
Security incidents	*Concerns about security incidents via technology or enabled by technology, such as misuse of related IT systems*	Winters, 2017; Woldeamanuel & Nguyen, 2018	4, 5, 13
Immaturity	*Concerns that technology is not yet fully mature and does not meet functional expectations*	-	2, 6, 7, 8
Data			
Privacy incidents	*Concerns about data privacy, in particular the use of and access to personal data (privacy invasion), disclosure of personal data to third parties (e.g., employers and health insurance companies), misuse of personal data for other purposes, and loss of control over the usage of personal data*	Alawadhi & Hussain, 2019; Coudert, 2010; Duarte, 2017; Newell & Marabelli, 2015; Strobel, 2019; Winters, 2017; Woldeamanuel & Nguyen, 2018	1, 6, 7, 8, 9, 10, 11, 12, 13
Data manipulation	*Concerns that manipulated data underlying the algorithm may lead to biased results of ADM*	Winters, 2017; Yang et al., 2018	1, 5, 9
Insufficient or wrong data basis	*Concerns that the data basis is insufficient, or that the data provided cannot be explained adequately*	-	3, 4, 6, 9, 10, 13
Decision			
Poor decision quality	*Concerns about the poor decision-making quality of a given system, leading to mistakes or decisions that do not match the fears, wishes, and preferences of individuals*	Bahner et al., 2008; Brauneis & Goodman, 2018; Strobel, 2019; Uhl, 1980; Westin et al., 2016; Winters, 2017; Woldeamanuel & Nguyen, 2018	1, 2, 3, 4, 7, 8, 11, 12, 13
Lack of transparency and missing verifiability	*Concerns about the lack of traceability of decisions by ADM, as decision-making takes place in the background ("black box") and is thus not comprehensible for individuals*	Brauneis & Goodman, 2018; Strobel, 2019; Westin et al., 2016; Yang et al., 2018	1, 3, 8, 9, 13
Fading of individual influence	*Concerns about losing the ability to influence the decision-making process due to loss of personal bargaining power, as opposed to traditional decision-making*	-	1, 2, 3, 6, 7, 11
Omission of human decision factors	*Concerns about the lack of human elements (empathic capacity) in ADM's decision-making, i.e., soft aspects and special cases are no longer taken into account*	-	1, 2, 5, 8, 9, 10, 12, 13

Table 3.1-3: Individuals' inherent concerns about ADM

The first category, technology, describes concerns about the technology used for ADM. Breakdown of technology is primarily seen as dangerous because “humans are highly dependent on technology” (Interviewee 10) and because technology could create “accidents involving humans” (Winters, 2017). The literature also shows that individuals are concerned about disruption to infrastructure (Winters, 2017) or potential system failure (Woldeamanuel & Nguyen, 2018). Security incidents refer to security concerns about system as a whole, and especially to the underlying data. Woldeamanuel and Nguyen (2018) indicate that the majority of individuals has security concerns, be they clear-cut security incidents or more general concerns about incidents associated with technology, e.g., the fear that someone may know when you are not home and then “burgle the house” (Interviewee 5, 13). Doubts that the system “will ever be mature enough to work 100%” (Interviewee 7) are summarized in the category immaturity.

The category data comprises concerns that individuals expressed about data used for ADM. Privacy incidents facilitated by “complete transparency of individuals” (Interviewee 1) are widely discussed in the literature (Alawadhi & Hussain, 2019; Coudert, 2010; Duarte, 2017; Newell & Marabelli, 2015; Strobel, 2019; Winters, 2017), and indeed in our interviews. The statements of Interviewee 12 (“the idea that one is completely predictable is daunting”), Interviewee 13 (who expressed concern about “having no control at all” over personal data), or Interviewee 11 (who said “data collected will be used for any other purpose”) confirm the relevance of this issue. Concerns about manipulation of “the input that the algorithm receives” (Interviewee 9), e.g., via “false statements” (Interviewee 1), “paid advertisement that influences the algorithm” (Interviewee 1, 5), or that “small changes in the input data […] may lead to drastic changes in the output, making the result uninformative and easy to manipulate” (Yang et al., 2018) are summarized in data manipulation. Insufficient or wrong data basis includes, e.g., concerns about “weak points in the entered data, where you know they can be misinterpreted without further explanation” (Interviewee 3) or that data quality “depends on how well I maintain my personal data, e.g., how I answer the questions” (Interviewee 13), related to the thought that “an algorithm needs all data from my wife and me, and so it is not capable of booking a holiday for us, as it will never know how many and which compromises are possible and which are not” (Interviewee 10).

The category decision presents concerns that individuals have about the automated decision itself. Individuals are concerned about poor decision quality. They are convinced that “implementation will never be 100% correct” (Interviewee 1) and think that the algorithm cannot respond with sufficient sensitivity to highly individual needs. The topic of poor decision quality is also discussed in the literature as the fear individuals have, for example, about incorrect decisions (Strobel, 2019) or false recommendations. Furthermore, individuals are concerned about a lack of transparency and missing verifiability of decisions made by ADM, i.e., they cannot verify whether the decision really is the best one or "if it is only the third-best offer" (Interviewee 1), because they "don't know about the decision basis in the background" (Interviewee 1). Intransparency is another relevant topic in the literature, as individuals do not fully understand the opacity of a system (Westin et al., 2016). Meanwhile, fading of individual influence is discussed in the interviews with regard to "loss of bargaining space" (Interviewee 1) or a sense that there is no "possibility for a personal introduction, where my abilities might be recognized" (Interviewee 11) due to a lack of human involvement. Omission of human decision factors refers to "missing empathy" (Interviewee 9), "complete reduction to numbers" (Interviewee 5), and the thought that a "human can be better assessed by other humans than by algorithm" (Interviewee 12), especially in "exception cases" (Interviewee 12).

Having illustrated the inherent concerns, Table 3.1-4 shows concerns about consequences of ADM.

Concerns	**Description**	**Literature sources**	**Interviews**
Physical			
Physical harms	*Concerns that use of ADM may result in physical harm, such as accidents involving individuals*	Brauneis & Goodman, 2018	6, 7, 13
Psychological			
Psychological harms	*Concerns that the feeling of being at the mercy of ADM systems has negative consequences on individuals' mental health*	-	13
Social			
Discrimination	*Concerns that existing discrimination in human decision-making is being systematized through ADM, leading to structural biases and unfairness in decisions*	Albarghouthi & Vinitsky, 2019; Binns et al., 2018; Brauneis & Goodman, 2018; Dietrich & Weisswange, 2019; Kullmann, 2018; Persson & Kavathatzopoulos, 2017; Strobel, 2019; Veale & Edwards, 2018; Woldeamanuel & Nguyen, 2018; Yang et al., 2018	1, 2, 3, 4, 6, 7, 8, 11, 13
Resource-related			
Negative financial impact	*Concerns about ADM making decisions that are financially unfavorable for individuals*	-	6, 8, 13

Environmental harms	*Concerns about negative impacts on environment through the spread of ADM*	Winters, 2017; Woldeamanuel & Nguyen, 2018	-
Prosecution-related			
Obscure legal regulation of responsibility	*Concerns about missing or unclear legal accountability for the decisions taken by algorithms*	Binns et al., 2018; Persson & Kavathatzopoulos, 2017; Woldeamanuel & Nguyen, 2018	2, 3
Career-related			
Job loss	*Concerns about becoming unemployed due to widespread use of ADM*	Winters, 2017	-
Freedom-related			
Monopolization of economy	*Concerns about monopolization on a limited number of platforms which gain disproportionate power from data, leading to a centralized and unbalanced market*	-	1, 8, 11
Skill loss	*Concerns about individuals losing abilities or skills because they are no longer used to performing certain tasks*	Winters, 2017	2, 7, 8, 10, 12
Obscure explicitation of value system	*Concerns about a lack of morality in ADM or a mismatch between the moral values of the system and personal values*	-	1, 2, 4, 8, 12, 13
Negative effects on human well-being			
External determination	*Concerns that individuals give up more control over their lives to ADM systems (and organizations operating those systems)*	Newell & Marabelli, 2015; Woldeamanuel & Nguyen, 2018	3, 4, 6, 8, 9, 11, 12
Lack of enjoyment	*Concerns that ADM decreases sensual and joyful moments, as the decision-making process itself is an enjoyable part of life that is no longer experienced by humans*	-	2, 6, 8, 9, 10, 13
Lack of individuality	*Concerns that ADM is not capable of reaching a level of individuality close to that of highly individual human decision-making*	-	1, 2, 8, 9, 12, 13
Lack of spontaneity	*Concerns that rigid patterns of ADM curtail the human value of spontaneity in daily life*	-	1, 4, 7, 13

Table 3.1-4: Individuals' concerns about consequences of ADM

Physical harms refer, for the most part, to accidents caused by ADM, e.g., via self-driving cars and other health hazards due to the increasing use of ADM technologies. In contrast, *psychological harms* denote "emotional damage" (Interviewee 13) through ADM. In the literature, Brauneis and Goodman (2018) also mention the concern that data can be used to hurt individuals.

Discrimination is among the most frequently discussed topics in the literature on ADM. Perhaps the most common form this takes is gender discrimination against individuals or protected groups (Kullmann, 2018; Persson & Kavathatzopoulos, 2017; Yang et al., 2018). Our interviews confirm this, as many interviewees fear biased decisions due to the "discrimination between men and women" (Interviewee 8) and "exclusion of people who cannot afford or use technologies that get more and more sophisticated and therefore expensive" (Interviewee 11). Often, discriminatory decisions made by automated systems result from biased training data sets (Interviewee 1).

Furthermore, individuals are concerned about ADM having a *negative financial impact*, primarily caused by *data manipulation*, e.g., when an algorithm orders a product at "a disadvantageous price" due to a paid advertisement (Interviewee 8) or a faulty product that will not be used (Interviewee 6). The category *environmental harms* comprises aspects such as increasing air pollution or greenhouse gas emission (Winters, 2017; Woldeamanuel & Nguyen, 2018).

The following concern *obscure legal regulation of responsibility* is prosecution-related. Individuals fear that it is unclear "who bears responsibility if something happens" (Interviewee 3). One such concern relates to the use case of autonomous driving, as stated by Interviewee 2: "In case somebody dies, or gets injured or anything else, who is responsible?"

Individuals also have career-related concerns. Winters (2017) states that individuals fear losing their jobs (*job loss*) due to ADM.

The first concern in the category of freedom-related concerns is the *monopolization of economy*, meaning that "the market becomes more unbalanced" (Interviewee 1). *Skill loss* refers to the concern that with an increasing number of automated decisions and thus a diminishing proportion of human-made decisions, individuals lose human abilities, such as "empathy" (Interviewee 10) and decision-making skills (Interviewee 12). *Skill loss* also includes a concern about "humans becoming lazy or less industrious" and "losing certain abilities or skills" (Winters, 2017). In *obscure explicitation of value system*, individuals fear a lack of morality in ADM or a mismatch between the moral values of the system and personal values. For instance, this may result from distinct cultural backgrounds of an algorithm's programmer and its users.

The subcategory of negative effects on human well-being comprises four concerns. Individuals prefer non-binding "recommender systems" (Interviewee 2, 8) in contrast to a completely automated decision in order to avoid *external determination*. The literature confirms these views, as concerns about dependence and loss of control have already been investigated (Newell

& Marabelli, 2015; Woldeamanuel & Nguyen, 2018). *Lack of enjoyment* includes statements that ADM in private life is associated with having less fun. For example, decisions about food or traveling are perceived as “fun” (Interviewee 6, 9), and to some the decision-making process itself constitutes an “experience” (Interviewee 13), which is why some do not want to give up decision-making. Meanwhile, *lack of individuality* denotes concerns about the inability of ADM to reach a sufficiently high level of individuality in decision-making: “No matter how complex the algorithm, it will never offer a highly individual trip for me” (Interviewee 8). Another interviewee raised the question: “Where is the individuality?” (Interviewee 2). The *omission of human decision factors* is seen as the chief reason why ADM will not achieve sufficiently high individuality. Moreover, individuals are concerned about a *lack of spontaneity* through the use of ADM in their daily lives, as they feel that the algorithm cannot respond unprompted to changes, which is why there will no longer be any room for spontaneity (Interviewee 1). Incidentally, according to some individuals it is simply “nice if not everything is planned, but you just happen to stumble over something” (Interviewee 8).

3.1.5 Discussion

The interviewees confirmed concerns that were identified by the structured literature search. Only two concerns originating from the literature could not be confirmed by our qualitative content analysis (*job loss, environmental harms*). This might be due to the abstract nature of these two long-term consequences of ADM, which is to say that our interviewees may well have thought of those aspects as being too far in the future to be caused by single automated decisions. Yet these two aspects aside, the concerns discussed in the literature were supplemented by eleven further concerns that were first identified in our qualitative content analysis. To break down those numbers, four concerns were added to the literature on the left-hand side of the framework (immaturity, insufficient or wrong data basis, fading of individual influence, omission of human decision factors).

A closer look at the inherent concerns in Table 3.1-3 shows that only concerns in the category *decision* are unique to ADM. Conversely, *technology* and *data* concerns can also be transferred to other new technologies, such as the Internet of Things (IoT) or Blockchain. For example, *security* and *privacy incidents* have already been discussed in depth in the existing IoT literature (Leloglu, 2017; Naeini et al., 2005). Further concerns, such as *immaturity,* also pertain to other new technologies and are, therefore, not specific to ADM (Lepekhin et al., 2019). Concerns arising from these two categories – *technology* and *data* – can lead to concerns about consequences for individuals, organizations, or society, and these concerns can be held regardless of whether a specific automated decision is executed. For example, a *security incident* where personal data is stolen, which causes a *privacy incident*, might lead to discrimination in another context, one that is quite distinct from the original decision-making process during which the data was collected and therefore not governed nor indeed controlled by the initial decision.

As explained above, our framework contains eleven concerns that emerged solely from our interviews and have not been addressed in previous research. Within all categories, the interviews revealed new inherent concerns as well as concerns about consequences that lend themselves to further examination in future research, which is strongly recommended in order to reduce individuals' skepticism about ADM and improve its acceptance among users. Some of the associated concerns worthy of further research are as follows: first, interviewees mentioned several aspects that mitigate their concerns about ADM, chief among them the fact that for many there is no perceived difference between ADM and a human decision-making process. For instance, interviewees often do not see a notable difference whether they provide their personal data to a human or to an algorithm. Furthermore, they tend to think that nowadays many organizational processes are already automated to a high degree, even though a human employee is involved. Another crucial aspect that would seem to attenuate many concerns is transparency. If individuals think they understand the decision-making process, which is to say that if they understand how and why the algorithm comes to its decision, many concerns are mitigated. A research area that focuses on this phenomenon is called explainable AI (XAI). XAI research analyses the black-box problem, i.e., that AI is becoming ever more complex. Hence, it becomes more difficult for the user to truly understand how the system works, and this diminishes the transparency of the user system (Bahdanau et al., 2017). This, in turn, brings us to trust, the third mitigating aspect mentioned in our interviews. Individuals state that their concerns about a specific ADM system significantly decrease when they trust the system, for instance, if they have had good experiences with the same system in the past.

In addition to those attenuating aspects, interviewees mentioned potential positive aspects of ADM, as opposed to human-made decisions. These include time savings, less effort for individuals, less subjectivity and more fairness in decisions, variety and positive surprises through ADM, and lower error rate in decisions. Future research could be of interest to examine the relationship between those attenuating and positive aspects of ADM on the one hand and the afore-mentioned concerns on the other. It might be very helpful for the development of ADM systems to know which concerns could be addressed by which attenuating aspects and under which circumstance, or for instance in which use case a user will focus more on positive aspects and less on concerns.

To develop these findings into a coherent theory, we follow in the footsteps of Urquhart et al. (2010): “Theoretical integration means relating the theory to other theories in the same or similar field.” Since there is, at the time of writing this, no relevant theory to draw on with regard to ADM, we employ a related theory from the field of information privacy research. Specifically, we compare our framework with Karwatzki et al. (2017), who investigate adverse consequences of access to individuals’ information. What makes this comparison especially apt is that Karwatzki et al. (2017) examine individuals’ technology-related concerns and develop a comprehensive conceptualization and categorization in terms of physical, social, resource-related, psychological, prosecution-related, career-related, and freedom-related adverse consequences. In our own research, we transfer this categorization to the field of ADM and use it to structure individuals’ concerns about consequences, i.e., the right-hand side of our framework (see Table 3.1-4). What is more, we identify inherent concerns about technology, data or decisions, i.e., the left-hand side of our framework (see Table 3.1-3). Karwatzki et al. (2017) present very detailed manifestations in each category, i.e., concrete concerns (e.g., kidnapping and imprisonment, slander and bullying, stalking), and these also apply to ADM. For instance, the manifestation “financial loss (direct or indirect)” is very similar to our concern *negative financial impact* (Karwatzki et al., 2017). Another example is the manifestation “being fired”. This relates to our concern *job loss* (Karwatzki et al., 2017). However, Karwatzki et al. (2017) identified other manifestations, such as “time loss”, which do not apply to ADM as they are mentioned neither in the literature nor in our interviews.

Moreover, we expect our framework to provide several meaningful insights for individuals and organizations using ADM, and it is our express hope that our work in this area will lead to further research. ADM is a current topic of great interest and potential, but so far researchers have focused either on the possibilities of using and implementing ADM or on dealing with its technical consequences and ethical issues, while the concerns of individuals have only been considered selectively or disregarded entirely. None of the papers to date have focused on any reasons for reluctance from an individual’s point of view. Our primary theoretical contribution is, therefore, the understanding and structuring of concerns that prevent individuals from using ADM applications. Our framework can be used – either ex-ante or ex-post – to anticipate and evaluate problems associated with the introduction of ADM applications. We believe that our framework provides an interesting new perspective on this issue and will guide future research.

Furthermore, it contributes to the extensive literature on the dark side of IS since it contains individuals' concerns and fears about using a specific technology, i.e., ADM.

Our results also offer practical benefits. A thorough consideration of concerns is essential as it can determine whether or not ADM applications are successfully disseminated. Our findings clearly show that some of the concerns are subjective feelings. Companies that implement or think about implementing ADM use cases should consider these concerns when developing ADM applications. They can use the framework to address these concerns, offer their prospective users targeted information, and strengthen trust in process outcomes based on automated decisions. Furthermore, our framework allows individuals to systematically gather information about ADM's potential risks for themselves and thus balance their concerns about ADM applications with facts. Many interviewees did not raise many concerns at the beginning of our interview but instead required concrete use cases to articulate their concerns.

Nevertheless, our research does not yet go far enough. Whereas the findings from the literature review are based on studies from different regions and countries, the interviews were all conducted in Germany. Expecting interesting cultural differences, the framework may be improved by extending the scope of the interviews to different countries (Belanger & Crossler, 2011). Even though we included open questions regarding concerns about ADM at the beginning of each interview, future research may strive for more generalizability or test concerns for a specific use case. Moreover, future research may clarify the relationship between the concerns by collecting quantitative data and evaluating it, e.g., with factor analysis. Such future research may also contribute to the current discussion by developing appropriate countermeasures that address individuals' concerns about ADM.

3.1.6 Conclusion

The aim of this paper was to provide an overview of concerns about ADM and thus show the need for further research in this area. To date, the literature in the field has neglected the individual human side. Therefore, it has failed to account for the importance of individuals' concerns as limiting factors in the adoption of ADM. Based on a thorough structured literature search and semi-structured interviews, we identified the concerns already addressed in the literature as well as those it has so far neglected. In total, we identified 24 concerns associated with integrating automated decisions into a person's life. We structured these concerns in a framework divided into different categories: technology, data, decision for inherent concerns, and concerns adapted from the categories of Karwatzki et al. (2017). It is our belief that this framework will help in summarizing and communicating concerns about ADM with a view to increasing confidence in automated decisions. As a result, this framework shall also support the adoption of ADM applications and enable individuals to be better informed about potential risks.

References

Alawadhi, R., & Hussain, T. (2019). A method toward privacy protection in context-aware environment. *Procedia Computer Science*, 151, 659–666. https://doi.org/10.1016/j.procs.2019.04.088

Albarghouthi, A., & Vinitsky, S. (2019). Fairness-aware programming. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA.

Allen, R., & Masters, D. (2019). Artificial intelligence: The right to protection from discrimination caused by algorithms, machine learning and automated decision-making. *ERA Forum*. Advance online publication. https://doi.org/10.1007/s12027-019-00582-w

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: there’s software used across the country to predict future criminals. And it’s biased against blacks. *ProPublica.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bahdanau, D., Cho, K., & Bengio, Y. (2017). Neural machine translation by jointly learning to align and translate. *IJCAI-17 Workshop on Explainable AI (XAI)*, 8(1).

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, 66(9), 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001

Belanger, F., & Crossler, R. E. (2011). Privacy in the digital age: a review of information privacy research in information systems. *MIS Quarterly*, 35(4), 1017–1041.

Binns, R., van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). It's reducing a human being to a percentage. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC, Canada.

Borràs, J., Moreno, A., & Valls, A. (2014). Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, 41(16), 7370–7389. https://doi.org/10.1016/j.eswa.2014.06.007

Boyatzis, R. E. (1988). *Transforming Qualitative Information: Thematic Analysis and Code Development*. SAGE Publications.

Brauneis, R., & Goodman, E. P. (2018). Algorithmic transparency for the smart city. *Yale Journal of Law and Technology*, 20(1).

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research,* 56(5), 809–825. https://doi.org/10.1177/0022243719851788

Cho, E., & Han, M. (2019). Ai powered book recommendation system. In D. Lo (Ed.), *Proceedings of the 2019 acm southeast conference on zzz - acm se '19* (pp. 230–232). ACM Press. https://doi.org/10.1145/3299815.3314465

Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S. (2016). *A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear*. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Coudert, F. (2010). When video cameras watch and screen: Privacy implications of pattern recognition technologies. *Computer Law & Security Review*, 26(4), 377–384. https://doi.org/10.1016/j.clsr.2010.03.007

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM,* 59(2), 56–62. https://doi.org/10.1145/2844110

Dietrich, M., & Weisswange, T. H. (2019). Distributive justice as an ethical principle for autonomous vehicle behavior beyond hazard scenarios. *Ethics and Information Technology*, 21, 227–239. https://doi.org/10.1007/s10676-019-09504-3

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

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), 1-5. https://doi.org/10.1126/sciadv.aao5580

Duarte, N. (2017). Building ethical algorithms. *Scitech Lawyer*, 14(1), 32–37.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., . . . Williams, M. D. (2021). Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management,* 57, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Faliagka, E., Tsakalidis, A., & Tzimas, G. (2012). An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet Research*, 22(5), 551–568. https://doi.org/10.1108/10662241211271545

Gimpel, H., & Schmied, F. (2019). Risks and side effects of digitalization: A multi-level taxonomy of the adverse effects of using digital technologies and media. *Proceedings of the 27th European Conference on Information Systems* (ECIS2019), Stockholm-Uppsala, Sweden.

Hauff, S., Veit, D., & Tuunainen, V. K. (2015). Towards a taxonomy of perceived consequences of privacy-invasive practices. In *Proceedings of the 23rd European Conference on Information Systems* (ECIS), Münster, Germany.

Hsie, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15, 1277–1288.

Karwatzki, S., Trenz, M., Tuunainen, V. K., & Veit, D. (2017). Adverse consequences of access to individuals' information: An analysis of perceptions and the scope of organisational influence. *European Journal of Information Systems*, 26(6), 688–715. https://doi.org/10.1057/s41303-017-0064-z

Kim, W., Jeong, O.-R., Kim, C., & So, J. (2011). The dark side of the internet: Attacks, costs and responses. *Information Systems*, 36(3), 675–705. https://doi.org/10.1016/j.is.2010.11.003

Kullmann, M. (2018). Platform work, algorithmic decision-making, and eu gender equality law. *International Journal of Comparative Labour Law and Industrial Relations*, 34(1), 1–21.

Kumar, B. (2015). An encyclopedic overview of big data analytics. *International Journal of Applied Engineering Research*, 10(3), 5681–5705.

Leloglu, E. (2017). A review of security concerns in internet of things. *Journal of Computer and Communications*, 5(1), 121–136. https://doi.org/10.4236/jcc.2017.51010

Lepekhin, A., Borremans, A., Ilin, I., & Jantunen, S. (2019). A systematic mapping study on internet of things challenges. *IEEE*. 9–16 .https://doi.org/10.1109/SERP4IoT.2019.00009

Lowry, P. B., Cao, J., & Everard, A. (2011). Privacy concerns versus desire for interpersonal awareness in driving the use of self-disclosure technologies: The case of instant messaging in two cultures. *Journal of Management Information Systems*, 27(4), 163–200.

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.

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

Mayring, P. (2010). *Qualitative Inhaltsanalyse*. In G. Mey & K. Mruck (Eds.), Handbuch qualitative forschung in der psychologie (pp. 601–613). VS Verlag für Sozialwissenschaften.

Naeini, P. E., Bhagavatula, S., Habib, H., Degeling, M., Bauer, L., Cranor, L., & Sadeh, N. (2005). Privacy expectations and preferences in an iot world. In *4th USENIX Conference on File and Storage Technologies,* San Francisco, CA, USA.

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems,* 24(1), 3–14. https://doi.org/10.1016/j.jsis.2015.02.001

Persson, A., & Kavathatzopoulos, I. (2017). How to make decisions with algorithms: ethical decision-making using algorithms within predictive analytics. *ACM Computers & Society*, 47(4).

Pirkkalainen, H., & Salo, M. (2016). Two decades of the dark side in the information systems basket: Suggesting five areas for future research. In *Proceedings of the 24th European Conference on Information Systems (ECIS),* Istanbul, Turkey.

Ringe, A., Dalavi, M., Kabugade, S., & Mane, P. P. (2019). Iot based smart refrigerator using raspberry pi. *International Journal of Research and Analytical Reviews,* 154–158.

Rustemeyer, R. (1992). *Praktisch-methodische Schritte der Inhaltsanalyse*. Aschendorff.

Schreier, M. (2013). *Qualitative content analysis*. In U. Flick (Ed.), The sage handbook of qualitative data analysis. SAGE Publications Ltd.

Shankararaman, V., & Gottipati, S. (2015). A framework for embedding analytics in a business process. In D. Aveiro & A. Caetano (Eds.), 2015 *IEEE 17th conference on business informatics* (cbi) (pp. 49–54). IEEE. https://doi.org/10.1109/CBI.2015.10

Smith, H. J., Milberg, S. J., & Burke, S. J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly,* 20(2), 167–196.

Strobel, M. (2019). Aspects of transparency in machine learning: doctoral consortium. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems* (AAMAS 2019), Montreal.

Tarafdar, M., Gupta, A., & Turel, O. (2015a). Introduction to the special issue on ‘dark side of information technology use’ - part two. *Information Systems Journal*, 25(4), 315–317. https://doi.org/10.1111/isj.12076

Tarafdar, M., Gupta, A., & Turel, O. (2015b). Special issue on ‘dark side of information technology use’: An introduction and a framework for research. *Information Systems Journal*, 25(3), 161–170. https://doi.org/10.1111/isj.12070

Uhl, F. S. (1980). Automated capital investment decisions. *Management Accounting*, 61(10).

Urquhart, C., Lehmann, H., & Myers, M. D. (2010). Putting the ‘theory’ back into grounded theory: Guidelines for grounded theory studies in information systems. *Information Systems Journal*, 20(4), 357–381. https://doi.org/10.1111/j.1365-2575.2009.00328.x

Vahn, G.-Y. (2014). Business analytics in the age of big data. *Business Strategy Review,* 25(3), 8–9. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8616.2014.01083.x

van den Broek, E., Sergeeva, A., & Huysman, M. (2019). Hiring algorithms: An ethnography of fairness in practice. In *Proceedings of the 40th International Conference on Information Systems (ICIS),* Munich, Germany.

Veale, M., & Edwards, L. (2018). Clarity, surprises, and further questions in the article 29 working party draft guidance on automated decision-making and profiling. *Computer Law & Security Review*, 34(2), 398–404. https://doi.org/10.1016/j.clsr.2017.12.002

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly,* 26(2), 13–23.

Westin, C., Borst, C., & Hilburn, B. (2016). Automation transparency and personalized decision support: Air traffic controller interaction with a resolution advisory system. *IFACPapersOnLine,* 49(19), 201–206.

Winters, J. (2017). By the numbers: How much do we trust ai? *Mechanical Engineering*, 139(1), 26–27.

Woldeamanuel, M., & Nguyen, D. (2018). Perceived benefits and concerns of autonomous vehicles: An exploratory study of millennials’ sentiments of an emerging market. *Research in Transportation Economics*, 71, 44–53. https://doi.org/10.1016/j.retrec.2018.06.006

Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H. V., & Miklau, G. (2018). A nutritional label for rankings. In G. Das, C. Jermaine, & P. Bernstein (Chairs), *Proceedings of the 2018 International Conference on Management of Data*, Houston, TX, USA.

3.2 Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI)

Abstract:

Philosophical and sociological approaches in technology have increasingly shifted toward describing AI (artificial intelligence) systems as '(moral) agents,' while also attributing 'agency' to them. It is only in this way – so their principal argument goes – that the effects of technological components in a complex human-computer interaction can be understood sufficiently in phenomenological-descriptive and ethical-normative respects. By contrast, this article aims to demonstrate that an explanatory model only achieves a descriptively and normatively satisfactory result if the concepts of '(moral) agent' and '(moral) agency' are exclusively related to human agents. Initially, the division between symbolic and sub-symbolic AI, the black box character of (deep) machine learning, and the complex relationship network in the provision and application of machine learning are outlined. Next, the ontological and action-theoretical basic assumptions of an 'agency' attribution regarding both the current teleology-naturalism debate and the explanatory model of actor network theory are examined. On this basis, the technical-philosophical approaches of Luciano Floridi, Deborah G. Johnson, and Peter-Paul Verbeek will all be critically discussed. Despite their different approaches, they tend to fully integrate computational behavior into their concept of '(moral) agency.' By contrast, this essay recommends distinguishing conceptually between the different entities, causalities, and relationships in a human-computer interaction, arguing that this is the only way to do justice to both human responsibility and the moral significance and causality of computational behavior.

Keywords: moral agency, human-computer interaction, artificial intelligence, responsibility, technical philosophy

Authors: Alexis Fritz, Wiebke Brandt, Henner Gimpel, Sarah Bayer

Status: This article is published in De Ethica: A Journal of Philosophical, Theological and Applied Ethics, Volume 6, Issue 1, pp. 3-22 (2020).

3.2.1 Introduction: Exemplary harmful outcomes

Artifacts have played a substantial role in human activity since the first Paleolithic hand axes came into use. However, the emergence of an (ethical) discussion about which roles can be attributed to the people and artifacts involved in an action is only a consequence of the increasing penetration of artifacts carrying ‘artificial intelligence’ (AI) into our everyday lives.

Let us consider three examples of the potentially harmful effect of sophisticated machine learning approaches:

1) Google’s search engine shows ads for high-paying executive jobs to men, but not so much to women (The Washington Post, 2015). Google’s photo tagging service incorrectly labeled photos showing African-American people as showing ‘gorillas’ (The Guardian, 2018b). Even years after being alerted to this racist behavior, Google did not fix the machine learning approach itself, instead simply removing the word ‘gorilla’ from the set of possible labels (The Guardian, 2018b).

2) Amazon developed a machine learning system designed to analyze the résumés of job applicants and rate them with respect to their technical skills. The system was shown to be sexist in how it distinguished between applicants: ‘It penalized résumés that included the word ‘women’s,’ as in ‘women’s chess club captain.’ And it downgraded graduates of two all-women’s colleges’ (The Guardian, 2018). Amazon eventually shut down the system after failing to fully prevent discrimination.

3) In pretrial, parole, and sentencing decisions in the U.S., machine learning algorithms frequently predict a criminal defendant’s likelihood of committing a future crime. The calculation of these so-called ‘recidivism scores’ is made by commercial providers that do not disclose the workings of their models. It was demonstrated for a widely used criminal risk assessment tool that used 137 features concerning an individual that the model performs no better than a simple logistic regression using just two features: age and the defendant's total number of previous convictions (Dressel & Farid, 2018). Yet, the seemingly more sophisticated 137-feature black box is being used in practice and has been accused of having a racial bias (Flores et al., 2016; The Washington Post, 2016).

We do not suggest that Google, Amazon, or the providers of criminal risk assessment tools are sexist, racist, or discriminatory by purpose in any other way. These examples merely illustrate that even well-intentioned initiatives using subsymbolic AI black boxes can lead to harmful outcomes. These systems may do very well with respect to some performance measures but may have inductive biases which are hard to detect and hard to fix. Overall, applications of AI, and especially subsymbolic machine learning-based systems, are part of complex socio-technical systems. There is no doubt that AI systems have moral impact, but do they act and reason morally? (The Washington Post, 2016)

The question of whether it is possible to create ethically acting machines represents an ongoing discussion (Anderson & Anderson, 2007; Crnkovic & Çürüklü, 2012). Additionally, the dominant approaches of technical philosophy and sociology currently emphasize the moral significance of AI systems, and have moved towards calling them '(moral) agents' and attributing them 'agency.' The principal argument of this approach is that it allows us to describe both the moral effect of an action's technological components and the complex network of human-computer interaction in a sufficiently descriptive and ethical manner. It is therefore crucial to elucidate the semantics of 'agency' and 'moral agency,' as well as their connection to the concept of responsibility, in order to provide more clarity in settings involving hybrid human-computer intelligence. The central issue is whether we can better grasp the descriptive and normative dimensions of AI and especially subsymbolic machine-learning-based systems with the help of the 'agency' attribution.

In the first part of this research, we provide basic information on symbolic and subsymbolic AI, the black box character of (deep) machine learning, and the complex relationship networks in the supply and application of machine learning.

The second part elaborates ontological and action-theoretical basic assumptions of agency attribution regarding the current teleology-naturalism debate, as well as an explanatory model of Actor-Network Theory (ANT).

Thirdly, three technical philosophical models describing computer systems as '(moral) agents' are critically analyzed with regard to whether an extended agency attribution really illuminates the descriptive and ethical-normative structure of human-computer interaction, or whether it obscures this.

3.2.2 Background on artificial intelligence

AI describes a computer 'system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation' (Kaplan & Haenlein, 2019). Different levels of AI include narrow AI (below human-level intelligence, outperforming humans in specific domains but not being potent in other domains), general AI (human-level intelligence across many domains), and artificial super intelligence (above human-level intelligence). Contemporary AI systems show narrow AI (also known as weak AI).

Early computer programs solved tasks that can logically be described with a set of rules and are therefore easy for computers but require prolonged effort for people. A branch of AI still follows this route: computers are equipped with a formal representation of knowledge about the world and the rules of logical reasoning. Thus, they deductively generate new insights. This type of AI is *symbolic AI* because it builds on explicit symbolic programming and inference algorithms. IBM's chess computer Deep Blue defeating the chess world champion Gary Kasparov in 1997 is an example of a symbolic (narrow) AI system. The other type of AI is *subsymbolic AI* using machine learning. The challenge for today’s computer programs is to solve tasks that for humans are hard to describe formally, as they are more intuitive; for example, speech recognition, face recognition, or emotions (Kaplan & Haenlein, 2019). Machine learning aims to build computers that automatically improve through experience (Russell & Norvig, 2016). A computer program learns from experience with respect to a class of tasks and a specific performance measure, if its performance on tasks of that class improves with experience (Mitchell, 1997). However, this focus on experience might lead to an inductive bias if training data is not representative of the data and situations a machine learning model will face after training. Within AI, ‘machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications.’ (Russell & Norvig, 2016, p. 255) Contemporary voice assistants, such as Amazon’s Alexa, Apple’s Siri, and Microsoft’s Cortana, leverage such subsymbolic (narrow) AI.

Symbolic AI is easier to debug, easier to explain, and easier to control than subsymbolic AI, as symbolic programming lends itself to human inspection. Subsymbolic AI requires less upfront knowledge, builds on learning from data more successfully and shows better performance than symbolic AI in many domains, especially on perceptual tasks.

Deep learning is a form of machine learning that has gained popularity in recent years due to advances in (big) data availability, (cloud-based) massive computing power, algorithms, and openly available libraries for using these algorithms. In this context, the 'depth' refers to the number of layers in the network’s structure; for example, in an artificial neural network (ANN). In the training phase, the strength of the connections (an analogy to brain synapses) between different nodes (an analogy to brain neurons) in the network is identified and learned. The more nodes and connections a network has, the better the network can acquire structural descriptions of the domain (if sufficient training data is available). Some of the largest artificial neural networks have millions of nodes and billions of connections.

Black box character of (deep) machine learning

Machine learning models, especially deep ANN, are frequently perceived as a black box (Castelvecchi, 2016). Once such a model is then trained, and calculating the output based on a given input is rather simple. In principle, all the weights and functions to apply can be inspected manually. However, the sheer number of nodes and connections in a deep ANN, as well as the non-linearity of the calculations, make it practically very difficult, if not impossible, to fully understand the model's behavior for all but the most trivial examples. It is even more difficult to ex-ante predict the outcome of the statistical learning process. Thus, many people effectively perceive deep learning as a black box.

Over recent years, applications of AI became more sophisticated in terms of high-impact and high-risk tasks, such as autonomous driving or medical diagnosis. This has led to an increasing need for explanations (Zhu, 2018). At the same time, this rising complexity has made it more difficult to get insights and to understand and trust the system's functions – not just for users, but also for the programmers of those algorithms (Mitchell, 1997). A logical model, like a decision-tree with statements involving 'and,' 'if-then,' etc., is comprehensible for the user. The larger the decision tree, the longer it takes, but humans are able to work through this process. Understanding deep learning models with millions or even billions of connections can be compared to understanding human predictions: we might anticipate what the system predicts, based on prior experience with the system, but we will never be completely sure if our assumption about the system's operating principles is correct.

This lack of transparency stands at the core of the discussion about the accountability and responsibility of humans regarding AI systems: can the user trust a prediction or be responsible for a decision made by a system that she or he cannot understand? To solve this issue, the research stream of *explainable AI* discusses two main options: white box and black box approaches. White box approaches aim at transparency, for instance, by displaying verbally or graphically the 'information contained in the knowledge base,' or via explaining the evidence, such as displaying the symptoms and test results that indicate the existence of a disease (Lacave & Díez, 2002). As the operating principles of linear models or decision trees are easier to understand, those models still dominate in many application areas (Montavon et al., 2018). Nevertheless, complex machine-learning models are in the fast lane and should offer explanations of their predictions to users. Due to the rising complexity of such systems, we cannot expect users to understand how the models work (Biran & McKeown, 2017).

Taking the example of an ANN, black box approaches focus on, for example, visualizing the input-output relationship, thus showing which input is most responsible for reaching a certain output (Fong & Vedaldi, 2017; Zhu, 2018). These approaches help users and programmers shed light on the black box, but they do not reveal the whole complex functions of the ANN. Therefore, such approaches make AI 'more of a grey than a black box' (Zhu, 2018). Still, these highly performant black and grey box machine learning systems pose challenges in terms of agency, especially as these artifacts are part of complex systems involving multiple actors.

Complex relationship networks in the supply and application of machine learning

Figure 3.2-1 is a stylized picture of the value chain from algorithm development, all the way through to the human being affected by a decision. It is an abstract depiction of the processes behind the examples given above. By showing the different types of human actors involved, it can thereby illustrate the complex interplay between different human actors and artifacts.

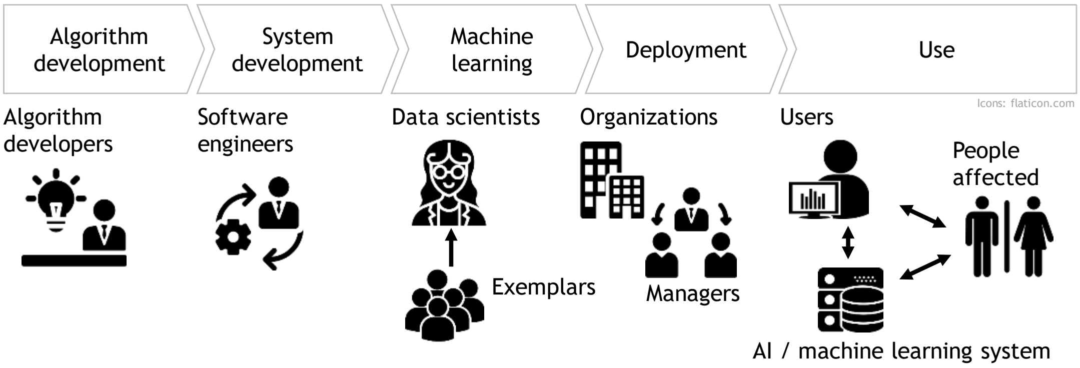


Figure 3.2-1: Stylized value chain from algorithm development to use of machine learning systems

Algorithm development conceives general-purpose machine learning algorithms. System development embeds these algorithms in a software system, typically for a specific purpose like criminal risk assessment or personnel decisions. The system is trained on the basis of data that originates from it (e.g., prior decisions by humans like evaluating résumés or sentencing criminals). Organizations like a court system or a company – or, more specifically, managers within an organization – then decide to use the system. Finally, individual users (like a clerk in the personnel department or a judge) interact with the machine learning-based system to obtain information and make decisions that affect others, like applicants or defendants.

If this overall socio-technical system harms people, who is responsible? There are eight candidates: (1) the technical AI system, despite it being an artifact; (2) the users obliged to use a system they do not understand; (3) the managers who neither understand the black box nor make individual decisions; (4) the organization; (5) the data scientists, despite the fact they do not make decisions concerning individual persons; (6) the people providing the training data, oftentimes unknowingly; (7) the software engineers, despite their inability to foresee the system's behavior after learning; and (8) the algorithm developers who created the multi-purpose black boxes in the first place. Is any single candidate responsible, several of them (each to a certain degree), is the overall socio-technical system responsible without individual responsibility, or are none of them responsible?

3.2.3 Pre-assumptions of agency attribution based on action theory

Asking what an actor or an action is and how it can be explained leads to a branched discussion of very different approaches to action theory. This makes it clear that agency attribution depends on several ontological and action-theoretical basic assumptions. Whoever uses concepts of action must not shy away from reflecting on these fundamental implications. Only against this background can different positions and their possible conclusions be adequately understood and discussed.

The teleology-naturalism debate concerns whether we can adequately describe and understand human actions and natural events by the same language and at the same level. Actor-Network Theory seeks to overcome the distinction between humans and non-humans by describing an actor as the symmetrical interplay between social, technical, and natural entities.

The teleology-naturalism debate in action theory

In order to determine the ways in which an action differs from a natural event, it is instructive to take a closer look at how we talk about it. We usually explain actions through the intentions of the person doing them ('She opened the window to air the room'), thus attributing the mental capacity to have goals, make decisions, etc. In contrast, we consider a natural event as the (provisional) end of a causal chain, and name the previous chain links as an explanation for its taking place ('The window opened because a gust of wind blew against it') (Runggaldier, 2010, p. 8). Obviously, we distinguish between a 'mental' language, which refers to actions, and a 'physical'(Runggaldier, 2010, p. 18) language, which refers to natural events (Runggaldier, 2010, p. 106). As long as both are applied only in their respective fields, there is no problem. However, it is questionable whether the same event can be expressed in both languages: is the window opening perhaps also due to certain neuronal states that triggered the woman's arm movement? Is such a physical description perhaps even more accurate than referring to mental states and abilities?

How do these different descriptions of the same event relate to each other? Are both of them legitimate perspectives that are able to coexist, or do they exclude each other so that at least one of them must be wrong? As a third option, one language might be translatable into the other (Sehon, 2010).

This is exactly the basic assumption of the naturalistic approach: anything expressed in mental language can be translated into physical language without any loss of meaning. Ultimately, there is no ontological difference between actions and natural events (Runggaldier, 2010). Accordingly, actions are subject to the same causal laws as natural events. Therefore, they can, in theory, be retrospectively deduced from a certain set of necessary and sufficient conditions, as well as predicted for the future if those very conditions are fulfilled (deductive-nomological explanatory scheme) – even if an accurate prediction is practically difficult to realize due to the complex interplay of numerous internal and external conditional factors (Runggaldier, 2010). In order to avoid this problem, a simpler action pattern is declared the object of investigation: the so-called 'basic action,' which consists of only a simple body movement (e.g. bending a finger) (Quitterer, 1998). If one regards the different levels of an action as an 'action tree,' then this 'basic action' represents the lowest, most basal level, which cannot be further explained by other partial actions. You get to higher levels by asking 'why?': he bent his finger to pull the trigger of a weapon, to fire a bullet at a person, to kill that person, etc. By contrast, you reach a lower level by asking 'how?': he killed him by shooting at him, by using the trigger, by bending the finger, etc. At this point, where you cannot break down the question of 'how?' any further, you have reached the lowest level (Runggaldier, 2010). Regardless of whether you consider these levels to describe the same action or many different actions,[14] both positions agree that the 'basic action' is the main, essential action on which further analysis has to concentrate.

The teleological approach contrasts with the naturalistic approach, and its followers criticize the orientation towards 'basic actions': in order to do justice to the nature of an action, it cannot be reduced to a body movement. On the contrary, the higher levels of the action tree are to be examined, where the actor‘s intentions, systems of rules and signs, the situational context with

[14] According to the 'unifiers'/'minimizers' bending the finger and killing the victim represent a single action; from the point of view of the 'multipliers'/'maximizers' these are numerically different actions (cf. Runggaldier, *Was sind Handlungen?*, pp. 50f; Quitterer, 'Basishandlungen', pp. 116f; Christian Budnik, 'Handlungsindividuation', in *Handbuch Handlungstheorie. Grundlagen, Kontexte, Perspektiven,* edited by Michael Kühler and Markus Rüther (Stuttgart: J. B. Metzler Verlag, 2016), pp. 60-68, at p. 60).

possibly involved third parties, etc. are situated (Runggaldier, 2010). Certain actions (e.g. greeting, betting, lecturing) are not dependent on a certain movement of the body, and therefore cannot be reduced to it (Runggaldier, 2010). But even actions whose correlation to body movements is evident, such as firing a weapon, are principally comprehensible only against the background of their circumstances and references: not the bending of the finger, but the intention to kill, the connection with the victim, etc., which constitute the action (Quitterer, 1998; Runggaldier, 2010). The reference to lower levels of action can be misleading, and even be used to deliberately conceal the essence of the action: 'I have only...' (Runggaldier, 2010).

Teleologists agree that intentions are the criterion that distinguishes an action from a natural event (Ricken, 2013). In contrast to the naturalistic translation thesis, they insist that mental language cannot be reduced to physical language, since intentions cannot be equated with the links of a causal chain (Horn & Löhrer, 2010; Runggaldier, 2010; Sehon, 2010).

Not only is it practically impossible to completely determine all the causal conditions for an action taking place, but this is also theoretically opposed by the conviction that a human being is fundamentally free in his decision to act (Runggaldier, 2010).

Donald Davidson, a representative of a moderate naturalism, takes this objection seriously and does not claim any principal predictability of human action. In the case of a broken windowpane, it can be stated afterwards, without any doubt, that a certain stone caused its breaking. However, to move from such a causal analysis to a prognosis about how hard one has to throw a stone against a window to break it in the future is something completely different. For actions, it applies analogously that individual, concrete actions can be explained causally and, in these individual cases, be translated into physical language. However, there are no laws either in the mental realm or between the mental and the physical sphere according to which predictions about future actions can be made. The name of this position, 'anomalous monism,' derives from the negation of such overarching laws.

Teleologists reply that such a concept devalues the mental side, since it is causally effective only insofar as it can be translated into physical terms (Runggaldier, 2010). Again, the intentionality of the actor is reduced.

Instead of searching for mental or physical events within the actor that have produced his action, one should simply accept the actor himself as the origin of his action ('agent-causality') (Runggaldier, 2010).

The concept of 'agency' in Actor-Network Theory (ANT)

Both naturalistic and teleological theories of action require a distinct separation between the subject and the object of an action. ANT criticizes this basic assumption. It opposes mechanistic, quasi-automatic explanations of actions, as well as models of understanding that presuppose the intention, autonomy, or consciousness of the human actor. But how are the terms ‘action’ and ‘agency’ to be understood if there is no subject-object difference, no primary principle, or no modern concept of the subject?

ANT is a challenging alternative to traditional theories of action, and has become one of the classic approaches of technical sociology (Häußling, 2019). Bruno Latour, Michel Callon, and John Law founded this theory in the 1980s and continue to develop it further to this day. Despite the diversity and complexity of the concepts within this family of ANTs, some key aspects shall be briefly highlighted.[15]

ANT does not ask why an actor acts in this way and not differently. Rather, it describes how an actor is transformed into an agent through the interplay of social, technical, and natural entities. The surprising thing is not so much that action always refers to others, but that non-humans are not simply passive objects of human action. Instead, they act themselves in a heterogeneous network (Latour, 1996).

This basic assumption is formulated by ANT as the general principle of symmetry, which claims a radically equal treatment of humans and non-humans. Social, technical, and natural factors are equal and depend on each other (Latour, 1995). In order to clarify the concept that not only humans are capable of acting, ANT replaces the ‘actor’ with an ‘actant.’ An actant is generally someone or something with the ability to act and to exercise activity (Akrich & Latour, 1992). Both human and non-human actants begin to create heterogeneous networks by themselves. They do not precede their networking but are produced by the networking process. The results of such networking are hybrids (i.e. hybrid forms of the social, the technical, and the natural) (Latour, 1995).

Actants transform into actors when a role and interests are assigned to them in the process of building networks (figuration) (Callon, 2006a). The successive and different steps of the network-building process are summarized under the term ‘translation.’ This is ‘the continuous attempt to integrate actors into a network by 'translating' them into roles and interests’ (Belliger & Krieger, 2006, p. 39). Translations create the ‘identities, characteristics, competences,

[15] A differentiated introduction to ANT in German is offered by Andréa Belliger and David J. Krieger, ‘Einführung in die Akteur-Netzwerk-Theorie’, in *ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie*, edited by Andréa Belliger and David J. Krieger (Bielefeld: transcript, 2006), pp. 13-50; Ingo Schulz-Schaeffer, *Sozialtheorie der Technik* (Frankfurt am Main: Campus-Verlag, 2000).

qualifications, behaviors, institutions, organizations and structures necessary to build a network of relatively stable, irreversible processes and procedures.' interests' (Belliger & Krieger, 2006, p. 39). A 'network' is not an external social reality, but a theoretical term for a concept that 'is traced by those translations in the scholars' accounts' (Latour, 2007, p.108). Statements about actants and actors are always moments in the process of network building or translation.

Latour exemplified his ANT by closing a door (J. Johnson, 1988). He understands this process as a network in which both human (= the user) and technical (= the door) actants are involved. If you regularly forget to close the door, this can quickly become a problem. This problem can then be solved, for instance, by introducing a sign, hiring a porter, or implementing a door-closing mechanism. If, for instance, a door-closing mechanism is installed, the new technical actant changes the characteristics and behavior of the existing network. For example, people have to adapt to the speed of the closing door.

While humans determine technical behavior, technical artifacts can also lead to human behavioral changes. In ANT, there is no clearly assignable making and being made; instead, there is only the network of actants (e.g. texts, people, animals, architectures, machines, or money) (Callon, 2006b).

This sometimes results in controversial, even irritating formulations in Latour's writing. Thus, a clumsy hotel key chain acts more morally than its human user. Due to its size, it forces the guest to hand in the key at the reception desk before leaving the hotel (Latour, 1991). When asked whether a person or a weapon was responsible for killing a person, Latour replied: 'It is neither people nor guns that kill. Responsibility for action must be shared among the various actants' (Latour, 1999, p. 180). It is a hybrid that cannot be reduced to a technical or human actant. Agency emerges from a connection of actants in the network: 'Action is a property of associated entities' (Latour, 1999, p. 182). Action and agency are always distributed among different entities. According to the sociologist M. Wieser, the notion of the agency in terms of non-human things must 'not be understood as animism or as the naive intentionality of things, but as the power of things, highlighting their resistance' (Wieser, 2012, p. 182). 'Agency' is not a substance, but a process (Wieser, 2012). In this sense, non-humans also possess the ability to act, for which the English term 'Agency' or 'Material Agency' has prevailed in technical sociology (Latour, 2007; Rammert, 2016; Wieser, 2012).

3.2.4 Three technical-philosophical approaches

It turned out that 'agent' or 'agency' are multifaceted concepts in the field of action theory. Their semantics and language practice depend on controversial and sometimes contradictory basic assumptions. The following technical-philosophical approaches are not identical with any of the action-theoretical directions discussed above. Nevertheless, the basic concerns, the course, or the focus of the following technical-philosophical approaches can each be traced back to one of the previously discussed theories of action.

The following approaches aim to describe and ethically evaluate the complex human-computer interaction appropriately and descriptively with the help of the terms '(moral) agent' or 'agency.'

The original problem and the basic concern of the three systemic models coincide. Nevertheless, Floridi's, Johnson's and Verbeek's answers compete with each other, and thus cannot be sensibly combined. To put it simply, we can describe Floridi's model as 'techno-centric,' Johnson's as 'anthropocentric,' and Verbeek's as 'constructivist.'

3.2.4.1 L. Floridi: Artificial agency

According to Floridi, the so-called standard ethics (i.e. deontological – like discourse-theoretical and contractualistic – or teleological – like virtue-ethical or consequentialist ethics) are hopelessly overwhelmed by the challenges of human-computer interaction (Floridi & Sanders, 2001). The first reason for this is that in conventional philosophy, only human beings (and thus no AI), are considered 'moral agents.' Thus, the human actor is burdened by a disproportionally great responsibility (Floridi & Sanders, 2004). Secondly, actions are judged on the basis of the actor's intentions (Floridi, 2016): it is morally relevant whether a person is injured intentionally or unintentionally. However, this focus on intentions does not help us where AI is used. In fact, the impact of a self-learning computer system can never be overlooked completely and therefore cannot be answered for by the designer or user. It is for this reason that Floridi suggests that we broaden the concept of 'moral agency' and refrain from judging intentions (Floridi, 2016).

Starting from the question who or what a 'moral agent' is, Floridi argues that definitions must be looked at in their particular context (Floridi & Sanders, 2004): A car mechanic looks at a car from a different point of view than an ethicist. To refer to these different points of view, Floridi uses the technical term 'level of abstraction.' At different levels of abstraction, different observables are relevant. For example, an ethicist delights in low pollutant emission, while a car mechanic is pleased by an unbroken V-belt (Floridi, 2010).

In order to define 'agent' properly, Floridi suggests a higher level of abstraction than is usually adopted. Candidates for 'agents' should no longer be examined for intentionality or other mental abilities; instead, they should be observed from a more distant perspective, appearing only vaguely as ‘systems.’ To be called ‘agents,’ systems have to be interactive, autonomous, and adaptive (Floridi & Sanders, 2004).

According to Floridi, whether, for example, a computer program checking CVs is considered an ‘agent’ depends on the granularity of the level of abstraction employed: if only the incoming CVs and their outgoing evaluation are regarded as ‘observables,’ but the algorithm itself is hidden, the recruitment program appears interactive, autonomous, and adaptive, consequently, as an ‘agent’: ‘interactive,’ because it begins to work in reaction to an external input; ‘autonomous,’ because it arranges the many applications automatically – as in a black box –; and ‘adaptive,’ because it learns on the basis of the data records (Floridi & Sanders, 2004, p. 362).

From ‘agent’ to ‘moral agent’ takes only a small step: for Floridi, all ‘agents’ whose actions have morally qualifiable consequences are ‘moral agents’ (Floridi & Sanders, 2004). Consequently, the recruitment program is not only an ‘agent,’ but also a ‘moral agent,’ because its selection is sexually discriminatory.

However, the program is not morally responsible for its consequences, as responsibility requires intention, but intention does not matter at the level of abstraction chosen for ‘agency.’ According to Floridi, ‘moral agents’ without intentions are not morally responsible for their actions but accountable (Floridi & Sanders, 2004). If artificial ‘moral agents’ cause damage – by analogy with sanctions on people – they can be modified, disconnected from the data network, or completely deleted or destroyed (Floridi & Sanders, 2004).

Floridi finally concludes that his understanding of ‘moral agency’ and ‘accountability’ sufficiently clarifies the ethical questions of human-computer interaction: ‘The great advantage is a better grasp of the moral discourse in non-human contexts’ (Floridi & Sanders, 2004, p. 376).

This positive self-evaluation of Floridi has to be questioned:

First, the AI debate is – according to Floridi – about attributing responsibility. If we stick to this assumption, we cannot see how the existence of non-responsible ‘moral agents’ can help in the search for a culprit.

Second, Floridi's reference to non-human ‘moral’ sources of good and evil of all kinds is nothing new in itself: a serious illness, a large avalanche, a chainsaw, a rabid dog, or falling roof tiles can all cause human suffering. However, despite the damage, we would never speak of a ‘moral’ avalanche, chainsaw, disease, dog, or tile.

By calling computer systems ‘moral,’ we can neither describe their mode of action better (causality), nor come closer to resolving moral issues (evaluation of an action or attribution of responsibility).

It can perhaps be said that the novelty of Floridi's approach lies not so much in qualifying the impacts of computer systems as ‘moral’ but in perceiving them as ‘agents’ at a certain level of abstraction. However, would that take us any further descriptively or normatively? This raises three thoughts: first, the necessity of making computer systems ‘accountable’ (i.e. that they have to be reprogrammed or even switched off if deficient) may be realized without there being any need of calling them ‘moral agents.’ While we may call our computer names when it does not do what we want it to, we do not do so because we seriously believe it will somehow impress our computer. Second, not all links in a causal chain need to be called ‘moral agents’ in order to become the object of ethical thought. Even in the standard ethics scolded by Floridi, a moral evaluation of an action or the attribution of responsibility is only possible after a precise and sufficient description of the causal connections. Third, it must also be criticized that if something goes wrong, at the level of abstraction favored by Floridi, the question of responsibility can no longer be posed for AI as a ‘moral agent,’ since Floridi abstracts from human intention, and computer systems are accountable but not morally responsible. In this way, ethically questionable incentive structures emerge, where the responsible party can be excused prematurely.

Thus, the impression is reinforced that the term ‘moral agents’ in Floridi's explanatory model contributes nothing toward gaining a better descriptive and normative understanding of human-computer interaction. It can thus be dismissed without consequences, since ‘moral agent’ or ‘moral agency’ is an empty concept if separated from responsibility.

3.2.4.2 D. G. Johnson: Triadic agency

Deborah Johnson struggles to find a happy medium between two extremes: one position undermines human responsibility to the extent that computer systems are referred to as ‘moral agents,’ and Johnson explicitly criticizes Floridi's approach. Representatives of the other position, on the other hand, misjudge the moral quality of machine behavior since they regard technology as extra-moral.

In the course of a larger searching movement, Johnson developed the so-called ‘Triadic agency’ model. According to Johnson, a state is caused neither by man nor by the computer system alone, but by a differentiated interaction. Basically, ‘agency’ means a ‘capability to act.’ Johnson distinguishes between three forms of agency:

(1) ‘causal agency’: things have a causal effect (Schlosser, 2015);

(2) ‘intentional agency’: people act intentionally; their intention causes the action (Schlosser, 2015);

(3) ‘triadic agency’: these forms of ‘agency’ relate to each other and are more than the sum of their individual parts. When people cooperate with computer systems, then:

> a. the user wants to achieve a certain goal – in our case the Amazon HR department wants an efficient and effective personnel selection –and delegates this task to the designers;
>
> b. the designer project team creates the recruitment program;
>
> c. with the help of this program the initial goal is achieved. (D. G. Johnson & Verdicchio, 2018)

In the ‘triadic agency’ model, responsibility is attributed only to those who are able to act intentionally. Since AI has no intention, it bears no responsibility for its causal effectiveness. Only humans can be ‘moral agents’ due to their intentional capacity. People therefore remain responsible, even if they delegate increasingly complex tasks to AI. In the search for the responsible person(s), it has to be asked in the direction of the designer or user until a person (or a group of persons) is found. However, an answer to the question of how much responsibility each person bears cannot be found without also considering the technological component.

By differentiating between three modes of action, Johnson first succeeds in maintaining the ontological difference between man and machine in terms of action theory. This differentiation is not essentialist, since it does not refer to fixed descriptive characteristics, but to certain abilities. Secondly, although only human beings can be responsible, their responsibility can only be clarified if all components of action are considered. Because of the descriptive and normative significance of machine behavior, Johnson does not want to renounce the agency attribution.

However, Johnson's inclusive use of the term ‘agency’ gives rise to misunderstandings and side scenes, since one term refers to human beings, computer systems, and human-computer interaction. Johnson strives to name the difference and interrelationship between man and computer systems, but she shrinks from taking the final step and continues to call computer systems ‘agents.’ Unlike Floridi’s use of the term, Johnson’s ‘agency’ is not meaningless but misleading. It would have been more beneficial to use different terms such as ‘factor,’ ‘cause,’ or ‘actor’ in order to emphasize the specific descriptive and normative contribution of computer systems.

3.2.4.3 P.-P. Verbeek: Hybrid agency

Peter-Paul Verbeek's ‘mediation theory’ is based on Don Ihde's postphenomenological approach and Bruno Latour's ‘actor-network theory’ (Verbeek, 2006, 2011). Verbeek emphasizes the joint causality of man and technology. Hence, technology actively mediates between human beings and their environment (Verbeek, 2006, 2014). It does so on two levels: hermeneutically, by influencing human perception of the world, and pragmatically, in partaking in human action (Verbeek, 2006).

Returning to our example of a recruitment program, the question of how the human resources department perceives the applicants – as deficient or positive – is decisively mediated by technology (hermeneutical mediation), and the final recruitment decision is pragmatically mediated. It is neither determined by, nor can it be made completely independently of, technology.

Consequently, according to Verbeek, moral decisions and actions are joint products of human beings and technology (Verbeek, 2014); morality is ‘hybrid,’ and ‘moral agency’ is a mixture (‘composite moral agency’) (Verbeek, 2014). No thing or living being possesses ‘moral agency’ by itself. Rather, ‘moral agency’ results from complex technical-human interaction; it does not form the basis for an action but emerges from it (Verbeek, 2014, 2017).

Verbeek goes so far as to describe even the actors themselves as the result of interaction (Verbeek, 2015). Nevertheless, Verbeek's theorem of a hybrid ‘moral agency’ does not mean that people cannot bear responsibility. In particular, designers of computer systems bear great responsibility because technology shapes the way of being in the world, and thus the human being himself. Verbeek shows the ethical dimensions with sentences such as ‘Designers materialize morality’ (Verbeek, 2015, p. 31) and ‘Designing technology is designing human beings’ (Verbeek, 2015, p. 28).

Against this background, we would like to ask whether Verbeek's ‘moral agency’ attribution helps us to understand human-computer interaction better both descriptively and ethical-normatively. The strength of Verbeek's postphenomenological-constructivist mediation theory undoubtedly lies in the fact that it acknowledges the complexity of human-computer interaction. Verbeek's approach is particularly successful in reflecting reality. If we accept that technology creates reality in terms of its interplay with human beings, and if this awareness replaces both obsession with, as well as forgetfulness about, technology, then much is gained for the debate about the responsible use of technology in both a descriptive and normative sense. This is true even if mediation is not a specific characteristic of technology alone.

However, with regard to Verbeek's understanding of ‘moral agency,’ there are important inquiries to make:

Unlike Floridi, Verbeek considers intentionality and freedom as part of the term ‘moral agency,’ albeit in a mediated, hybrid form. However, intentionality and freedom do not constitute ‘moral agency’. Instead, and much like ‘moral agency’ itself, this only results from a complex human-computer interaction.

The strength of the postphenomenological-constructivist view of reality turns into a weakness as soon as we want to attribute agency or responsibility to individual, concrete entities. In Verbeek's mediation theory, ‘moral agency,’ intention, freedom, and thus responsibility can no longer be attributed to individuals, since they always emerge from an overall structure. Ultimately, in Verbeek's theory of mediation, the individual and his actions cannot be conceived without technical influences or mediation. Human beings and computer systems are ‘actants’ – only as a mixture are they also ‘agents.’

Verbeek's two concerns – reconstructing the understanding of human-computer interaction and attributing moral responsibility – could also be fulfilled if the human actors remained ‘moral agents.’ For the realization that human capacity to act is always mediated is nothing new from a philosophical point of view. However, in order to avoid a circular conclusion in the attribution of ‘moral agency’ and moral responsibility, the freedom of human actors must be regarded as taking precedence. This is because interaction does not have its origin in itself but is a consequence of the human ability to reflect, decide, and act freely.

3.2.5 Conclusion

This study has revealed the opportunities and risks of applying the concept of ‘moral agency’ to human-computer interaction. Ultimately, the risks of agency attribution to computational behavior are disproportionate to the benefits of such language practice.

From a descriptive and ethical-normative point of view, this practice proves to be both unnecessary and risky. Floridi's use of ‘moral agents’ for computer systems is redundant.

Exclusive features for human or social contexts (e.g. ‘intentionality’ or ‘responsibility’), which should be preserved, come out of sight.

Verbeek offers a comprehensive and promising understanding of human-computer interaction. However, his ‘moral agent’ attribution is circular or leads to an infinite regression, thus making it objectionable. This is illustrated by the fact that it is difficult to identify a specific human capacity or actor for responsibility.

Johnson's results are consistent in view of their ontological and action-theoretical premises. She also conceptually differentiates the contribution of each component and is thus able to provide an almost accurate understanding of human-computer interaction. However, the ‘agency’ attribution gives rise to misunderstandings. At the same time, there is a serious risk that the extensive use of ‘moral agents’ undermines the question of responsibility.

Consequently, an appropriate differentiation between humans and computers should also be conceptually discernible. In this way, human-computer interaction can not only be described more precisely but the ethical-normative structure can also be elaborated more clearly.

References

Akrich, M., & Latour, B. (1992). *A summary of a convenient vocabulary for the semiotics of human and nonhuman assemblies. Shaping Technology/ Building Society*. Studies in Sociotechnical Change, Edited by Wiebe E. B. And John L. Cambridge, Mass.: The MIT Press.

Anderson, M., & Anderson, S. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4), 15–26.

Belliger, A., & Krieger, D. (2006). *Einführung in die akteur-netzwerk-theorie*. ANThology. Ein Einführendes Handbuch Zur Akteur-Netzwerk-Theorie, Edited by Andréa Belliger and David J. Krieger. Bielefeld: Transcript.

Biran, O., & McKeown, K. (2017). Human-centric justification of machine learning predictions. In *26th international joint conference on artificial intelligence*, Melbourne, Australia.

Callon, M. (2006a). *Einige elemente einer soziologie der übersetzung: Die domestikation der kammmuscheln und der fischer der s. Brieuc-bucht*. ANThology. Ein Einführendes Handbuch Zur Akteur-Netzwerk-Theorie, Edited by Andréa Belliger and David J. Krieger. Bielefeld: Transcript, 135–174.

Callon, M. (2006b). *Techno-ökonomische netzwerke und irreversibilität*. ANThology. Ein Einführendes Handbuch Zur Akteur-Netzwerk-Theorie, Edited by Andréa Belliger and David J. Krieger. Bielefeld: Transcript, 309–342.

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature*, 538(7623), 20–23.

Crnkovic, G. D., & Çürüklü, B. (2012). Robots: Ethical by design. *Ethics and Information Technology*, 14(1), 61–71.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1).

Flores, A., Bechtel, K., & Lowenkamp, C. (2016). False positives, false negatives, and false analyses: A rejoinder to ‘machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks. *Federal Probation Journal*, 80(2), 38–46.

Floridi, L. (2010). Levels of abstraction and the turing test. *Kybernetes*, 39, 423–440.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences,* 374(2083).

Floridi, L., & Sanders, W. (2001). Artificial evil and the foundation of computer ethics'. *Ethics and Information Technology,* 3, 55–66.

Floridi, L., & Sanders, W. (2004). On the morality of artificial agents. *Minds and Machines,* 14, 349–379.

Fong, R., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. *Proceedings of the IEEE International Conference on Computer Vision*, 3429–3437.

The Guardian. (2018a). *Amazon ditched ai recruiting tool that favored men for technical jobs*. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

The Guardian. (2018b). *Google's solution to accidental algorithmic racism: Ban gorillas.* https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people

Häußling, R. (2019). *Techniksoziologie. Eine Einführung*. Verlag Barbara Budrich.

Horn, C., & Löhrer, G. (Eds.). (2010). Ei*nleitung: Die Wiederentdeckung teleologischer Handlungserklärungen*. Suhrkamp.

Johnson, D. G., & Verdicchio, M. (2018). Ai, agency and responsibility: The vw fraud case and beyond. *AI & SOCIETY*.

Johnson, J. (1988). Mixing humans and nonhumans together: The sociology of a door-closer. *Social Problems*, 35(3), 298–310.

Kaplan, A., & Haenlein, M. (2019). Siri, siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons,* 62(1), 15–25.

Lacave, C., & Díez, F. (2002). A review of explanation methods for bayesian networks. *The Knowledge Engineering Review,* 17(2), 107–127.

Latour, B. (1991). *Technology is society made durable. A Sociology of Monsters? Essays on Power, Technology and Domination*, Edited by John Law. London/ New York: Routledge, 103–131.

Latour, B. (1995). *Wir sind nie modern gewesen. Versuch einer symmetrischen Anthropologie*. Akad.-Verlag.

Latour, B. (1996). *Social theory and the study of computerized work sites. Information Technology and Changes in Organizational Work*, Edited by W. J. Orlinokowsky and Geoff Walsham. London: Chapman and Hall, 295–307.

Latour, B. (1999). *Pandora's Hope. Essays on the Reality of Science Studies*. Harvard University Press.

Latour, B. (2007). *Reassembling the Social. An Introduction to Actor-Network-Theory*. Oxford University Press.

Mitchell, T. M. (1997). *Machine Learning*. WBC/McGraw-Hill.

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

Quitterer, J. (1998). *Basishandlungen und die naturalisierung von handlungserklärungen. Soziologische Handlungstheorie*. Einheit Oder Vielfalt. Edited by Andreas Balog and Manfred Gabriel. Opladen: Westdeutscher Verlag, 105–122.

Rammert, W. (2016). *Technik – Handeln – Wissen. Zu einer pragmatistischen Technik- und Sozialtheorie*. Springer.

Ricken, F. (2013). *Allgemeine Ethik*. W. Kohlhammer.

Runggaldier, E. (2010). *Was sind handlungen? Eine philosophische auseinandersetzung mit dem naturalismus*. In C. Horn & G. Löhrer (Eds.), Einleitung: Die wiederentdeckung teleologischer handlungserklärungen. Suhrkamp.

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

Schlosser, M. (2015). *Agency*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/agency/

Sehon, S. R. (2010). *Abweichende kausalketten und die irreduzibilität telologischer erklärungen.* In C. Horn & G. Löhrer (Eds.), Einleitung: Die wiederentdeckung teleologischer handlungserklärungen. Suhrkamp.

Verbeek, P.-P. (2006). *Materializing morality. Design ethics and technological mediation*. Science, Technology, & Human Values, 31, 361–380.

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

Verbeek, P.-P. (2014). *Some misunderstandings about the moral significance of technology. The Moral Status of Technical Artefacts,* Edited by Peter Kroes and Peter-Paul Verbeek. Dordrecht: Springer, 75–88.

Verbeek, P.-P. (2015). Beyond interaction: A short introduction to mediation theory. *Interactions*, 22, 26–31.

Verbeek, P.-P. (2017). *Designing the morality of things: The ethics of behaviour-guiding technology. Designing in Ethics*, Edited by Jeroen Van Den Hoven, Seumas Miller and Thomas Pogge. New York: Cambridge Univ. Press, 78–94.

The Washington Post. (2015). *Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you.* https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/

The Washington Post. (2016). *A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.* www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas

Wieser, M. (2012). *Das Netzwerk von Bruno Latour. Die Akteur-Netzwerk-Theorie zwischen Science & Technology Studies und poststrukturalistischer Soziologie.* transcript.

Zhu, J. e. a. (2018). Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. *IEEE Conference on Computational Intelligence and Games*, 1–8.

3.3 The role of domain expertise in trusting and following explainable AI decision support systems

Abstract:

Although the roots of artificial intelligence (AI) stretch back a few years, we are currently experiencing a flourishing of research and practical use. Already, AI has changed the world we live in, and it will continue to do so. However, AI deals with certain growing pains with a trust issue as the chief among them. To address this problem, science has favored the strategy of making AI explain itself to its user. So far, though, it is unclear how an AI can accomplish this in a way that increases trust and affects behavior in decision-making scenarios. This is especially difficult as users consume explanations differently, depending on their domain-specific expertise. With this in mind, this study focuses on how a user's expertise influences their trust in explainable AI (XAI) and how this, in turn, influences their behavior, i.e., their decisions. To test our theoretical assumptions, we develop an AI-based decision support system (DSS), observe user behavior in an online experiment, and complement it with survey data. The results show that domain-specific expertise negatively affects trust in AI-based DSS. We conclude that the strong focus on explanations might be overrated for users with low domain-specific expertise, whereas it is vital in generating trust in AI-based DSS among users with high expertise. Investigating the influence of expertise on explanations of an AI-based DSS, this study contributes to research on both XAI and DSS. Unlike most prior work in these areas, however, we go beyond assessing behavioral intention to investigate actual behavior.

Keywords: explainable artificial intelligence, trust, decision support systems, user expertise, online experiment

Authors: Sarah Bayer, Henner Gimpel, Moritz Markgraf

Status: This is an 'Accepted Manuscript' of an article accepted for publication by Taylor & Francis Group in Journal of Decision Systems, 2021,
https://www.tandfonline.com/toc/tjds20/current

3.3.1 Introduction

Artificial intelligence (AI) is literally everywhere – in our phones, our vehicles, our media consumption, and even in romantic matchmaking (Agrawal, 2018). Notwithstanding this apparent omnipresence, though, recent developments show that AI remains limited (Biran & McKeown, 2017; Miller, 2019). At first glance, this would seem to be a little odd, since those limitations are neither technical restrictions, nor are they caused by another technology outperforming AI. They are, instead, the result of a human limitation – our suspiciousness of the unknown, or rather of the things we do not understand. Our history has shown that we even demonize technical progress we fail to comprehend, like the railway or the telephone, because our ancestors who witnessed the emergence of such innovations were frightened of them (NRZ, 2017), or at least nervous enough to have reservations about using them. AI, maliciously nicknamed the "AI monster" (Ågerfalk, 2020), seems to fall into this category as a major limitation of AI in this day and age is its lack of use due to a lack of trust among the people for whom it was created (Miller, 2019; Ribeiro et al., 2016).

At present, AI has its most immediate impact on decision-making (Ågerfalk, 2020; Schmidt et al., 2020) via AI-based decision support systems (DSS). Those systems support the human user in a decision-making process by making suggestions, such as data-based predictions, or info summaries and displays. Nevertheless, the human determines the behavior as the final decision is up to the user, not the AI. After all, the role of AI-based DSS is to help humans make the best possible decisions. There are, however, famous examples of AI-based DSS that performed their role rather poorly. Examples include the AI security camera accusing an unequivocally innocent businesswoman of being a jaywalker (The Telegraph, 2018) or the Amazon AI recruiting system that favors men over women (Reuters, 2018). It stands to reason that such instances do not promote trust in AI-based DSS.

Trust plays a pivotal role in human relationships (Mishra & Morrissey, 1990) as well as human-computer interaction (Yan et al., 2011). When humans interact, it gives them an opportunity to get a feeling for the other person's benevolence, integrity, and competence (Mayer et al., 1995), which are the main building blocks of trust (McKnight et al., 1998). However, when one of the humans is replaced with an AI-based DSS, interaction is different, particularly because it is often one-sided as the user interacts with the system, but compared to a human, the system only responds in a limited fashion (W. Wang & Benbasat, 2008), and this creates a trust gap. One way to bridge or indeed close this gap is to make AI-based DSS explain their suggestions (Siau & Wang, 2018). Prior research, as summarized by Biran and Cotton (2017), provides evidence that this approach holds considerable promise because explanations increase user trust. Based on their findings, Biran and Cotton (2017) have called for more work to be done in this expanding field of research into trust in explainable systems. With this study, we are responding to this call.

The term explainable AI (XAI) denotes strategies to increase trust in AI-based systems by the use of explanations. XAI approaches fit into three different categories (Biran & McKeown, 2017): *visualization, interpretable models, and prediction interpretation & justification*. The category *visualization* covers methods such as highlighting cancer cells in color for magnetic resonance imaging (Lamy et al., 2019). Generally, visualization uses visual effects to explain, but for obvious reasons, such an approach is limited to visual problems. For further application scenarios, the XAI community distinguishes between the application of so-called white and black models. White models are inherently *interpretable models*. As such, they fall into the second category, meaning that these models use a training process that includes the creation of rules, decision lists, or decision trees, which humans tend to find easier to understand, at least to some extent. There is, however, a trade-off. While black models, such as artificial neural networks, usually outperform white models, their ready-trained models are far more difficult to understand (Lundberg & Lee, 2017). Approaches in the third XAI category, *prediction interpretation & justification*, mainly deal with those black models. Strategies here include investigating the importance of single attributes (e.g. Robnik-Sikonja & Kononenko, 2008) or transforming black models into white models (e.g. Ribeiro et al., 2016). Yet as well-performing models are becoming more complex and increasingly black, it is unreasonable to expect a user who is not an AI expert to understand AI models beyond their transparency or whiteness (Biran & McKeown, 2017). Users “expect explanations that they can understand” (Stahl et al., 2021, p. 384), and they expect explanations of AI performance to refer to the “same conceptual framework used to explain human behaviours” (Miller et al., 2017, p. 4). It is with this in mind that we conduct this study to investigate the XAI approach of justifying AI suggestions, rather than, for example, explaining AI functions. The extent to which explanations actually help to foster trust in an AI system is subject to current research. For instance, Schmidt et al. (2020) have found that plain transparency can harm trust. What characterizes a good explanation of an AI system remains an unresolved question (Miller, 2019), especially as a “good explanation has to be more than just true or likely to be true” (Hilton, 1996, p. 274). For quite some time now, the social sciences have tried to characterize good explanations. A recent study by Miller (2019) summarizes knowledge for the XAI research community. One of his key findings is that good explanations are social, meaning they are presented relative to the user’s beliefs or characteristics. This is advisable because different users absorb explanations differently, depending on their domain-specific expertise (Gregor & Benbasat, 1999). Since experts often have their own opinions predicated on their respective expertise, they tend to be rather surprised when facing a divergent opinion in an expert of a different discipline, but experts use explanations to resolve their disagreements (Gregor & Benbasat, 1999). In contrast, novices lack expertise, which makes them reliant on the opinions of third parties, and rather than question these opinions, they tend to use them to learn (Gregor & Benbasat, 1999). Expertise can, therefore, play an important role in designing AI and its explanations, which is why we aim to answer the following research question:

How does the domain-specific expertise of human users influence their trust in explainable AI decision support systems (XAI DSS), and does it affect their behavior to make them go along with the system's suggestions?

The answer to this question promises to deepen the understanding of how exactly explanations work in the area of AI. This, in turn, promises to reduce the trust issues surrounding AI and the related adverse behavior, which ultimately promises greater advances in AI and broader acceptance of the technology. To this end, we present our theoretical background and develop our hypotheses in section 2. This requires us to consider trust models based on the theory of reasoned action (TRA) (Fishbein & Ajzen, 1980) and its successor, the theory of planned behavior (TPB) (Ajzen, 1985). We also examine their application in the context of new technologies and contextualize them with regard to AI-based DSS and user expertise. In section 3, we describe our research process, which differs from most others dealing with trust in AI-based systems. Whereas they stop at the user's intention (e.g. Gefen et al., 2003; McKnight et al., 2002), we go one step further by performing an empirical examination of the user's behavior. More specifically, we conduct an online experiment via the crowdsourcing marketplace Amazon Mechanical Turk (mTurk). To choose a challenging cognitive task, we locate this experiment in the domain of chess. The game of chess is especially well-suited to test our theoretical hypotheses on XAI, user expertise, and trust, because it is the "most widely-studied domain in the history of artificial intelligence" (Silver et al., 2017, p. 1) and an excellent example of cognitive expertise (van der Maas & Wagenmakers, 2005). We then present our results in section 4, whereupon we discuss them in section 5 and conclude in section 6 with the integration in the scientific body of knowledge, where we also outline the limitations of our study and promising points of departure for further research.

3.3.2 Theoretical background and hypothesis development

Since we aim to examine trust in XAI DSS, we begin with a general overview of the theoretical background on AI and trust. Next, we look at the larger issue of trust in technology and, specifically, trust in AI. Against this background, we then derive our hypotheses, as depicted in Figure 3.3-1.

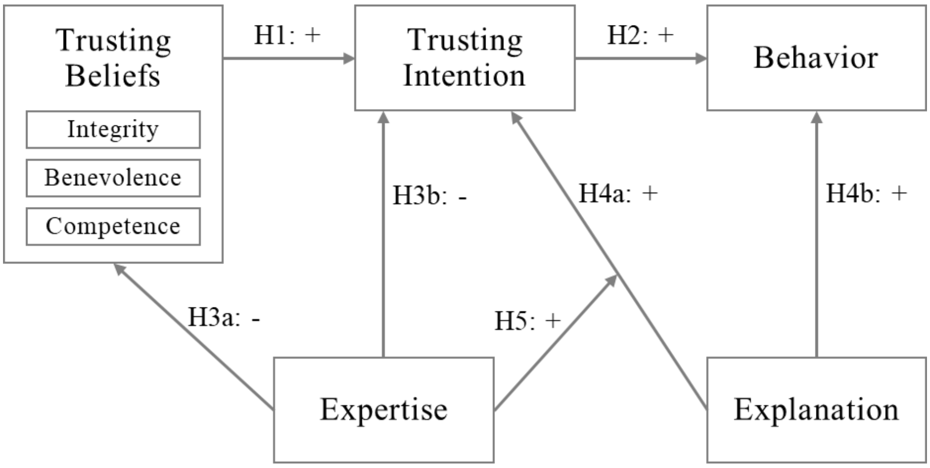


Figure 3.3-1: Research model

AI is a hot topic, both in science and in practice. Courting much debate and controversy, it has attracted various definitions and understandings (P. Wang, 2008). AI is associated with keywords like big data, analytics, neural networks, and machine learning (Ågerfalk, 2020; Fosso Wamba et al., 2020). The blurred boundaries between these concepts have even led to satirical send-ups, such as (Velloso, 2018): "Difference between machine learning and AI: If it is written in Python, it's probably machine learning. If it is written in PowerPoint, it's probably AI." Nonetheless, there are attempts at unification, including that of Rai et al. (Rai et al., 2018, p. iii): "AI is […] the ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem-solving, decision-making, and even demonstrating creativity." Another often-cited definition of AI is "the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment" (Russell & Norvig, 2016).

Independently of the preferred definition, AI systems can produce results in the shape of clustering or forecasts. We refer to these results as suggestions. What we are leaving aside, then, are more future-oriented, fully automatic AI-based processes, such as autonomous driving level 5, the kind of driving that does not require human attention or a steering wheel (SAE International, 2018). Instead, we focus on DSS to investigate the perceived trust in AI systems, since the impact of AI on human decision-making is more immediate (Ågerfalk, 2020) and the interactions between humans and AI are more substantial in the context of DSS (Power, 2002). The definition of DSS we subscribe to here posits them to be "interactive computer-based systems that help people use computer communications, data, documents, knowledge, and models to solve problems and make decisions." (Power, 2002, p. xii). Examples of such AI-based DSS include systems that provide advice for a medical doctor to find the correct treatment for a patient and systems that support the maintenance of the city's electrical grid (Bussone et al., 2015; Rudin et al., 2012).

Since such systems are entrusted with crucial tasks, users' trust undoubtedly plays a key role. In the scientific community, trust is a much-discussed concept, yet there is no consensus on its definition since it varies depending on its conceptualization and the researcher's background (Gefen et al., 2003; McKnight et al., 2002). Söllner et al. (2016, p. 1), for instance, define trust broadly as the "willingness of one party (the trustor) to rely on another party (the trustee)." Mayer et al. (1995, p. 712) define it more specifically as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party." However, there is widespread agreement that trust is multi-dimensional and multi-faceted depending on one's point of view (McKnight et al., 2002). In the following, we will examine this in greater detail with the AI-based DSS in the role of the trustee and the human user in that of the trustor.

Even before one has any experience with a trustee, there can still be trust in the form of *initial trust*, its main components being institutional and personality-based trust (Kim et al., 2009; Li et al., 2008; McKnight et al., 2002; Pavlou & Gefen, 2004) (or *trust credit* (Gefen et al., 2003)). With no direct experience, the trustor builds trust upon their intuition and personality (McKnight et al., 2002). Once experience is gained, this initial trust is gradually replaced by *knowledge-based trust* (Gefen et al., 2003). By the first experience via second-hand information or a first glance at the beginning of a relationship, one can assess the trustee's *integrity* (Gefen et al., 2008; Mayer et al., 1995). This denotes the belief as to whether or not the trustee is honest and keeps their promises (McKnight et al., 2002). As one becomes more experienced with the trustee, the trustor can also judge the latter's *benevolence and competence* (Gefen et al., 2008; McKnight et al., 2002). Benevolence means that the trustee's actions are advantageous to the trustor and done with righteous intention. Competence means that the trustee has the required skillset in the respective domain (McKnight et al., 2002). Along with integrity and benevolence, competence (otherwise referred to as *ability* (Siau & Wang, 2018; Toreini et al., 2020)) forms the extensively studied *trusting beliefs* (McKnight et al., 2002) or *ABI framework* (Toreini et al., 2020), and indeed its successor ABI+ (Dietz & Gillespie, 2011; Dietz & Hartog, 2006). Previous studies, such as McKnight et al. (2002), Gefen et al. (2003), Li et al. (2008), Wang and Benbasat (2005), have successfully applied them to assess trust in technologies. They would, therefore, seem to be applicable attributes for technologies and especially for those “designed and operated by humans” (Li et al., 2008, p.48). As we intend to examine the trust that participants develop during interaction with the AI system, we focus on knowledge-based trust.

To estimate behavior based on trusting beliefs, trust research in technological domains (e.g. Gefen et al. 2003, Li et al. 2008, McKnight et al. 2002) is most often based on TRA by Fishbein and Ajzen (1980) and its development, the so-called theory of planned behavior (TPB) (Ajzen, 1991). Both models suggest that an attitude towards behavior leads to an intention which, in turn, leads to behavior. However, the models differ in what they treat as influencing factors of attitude. With special regard to AI, previous trust research studied three different perspectives: First, a global perspective on trust in AI sees human, environmental, and technological characteristics as being essential (Siau & Wang, 2018). While these are universally valid characteristics, within the study of AI, a special focus lies on the technological characteristics as AI differs markedly from other technologies. The second perspective corroborates this as it focuses on the trustworthy development and design of AI (Toreini et al., 2020). For the third perspective – AI-based automation which replaces human experts with an AI system (Ribeiro et al., 2016) – trust research addresses the technology itself along with the associated organization (Hengstler et al., 2016). To the best of our knowledge, though, no one has yet developed a theory on trust in AI and trust-based behavior when interacting with DSS.

According to TRA, trusting beliefs positively affect trusting attitude, which denotes the “personal judgment that performing these behaviors would result in good or bad consequences” (Li et al. 2008, p. 48). The trusting attitude, in turn, positively affects the trusting intention, which denotes the intention to behave according to the trust (Ribeiro et al., 2016; Silver et al., 2017; Urbach et al., 2010). TRA and TPB form the foundation of trust models in the context of technology, such as the Technology Acceptance Model (TAM) of Davis (1989), and its successors, such as the often-cited trust models by Gefen et al. (2003) and Wang and Benbasat (2005). All of these models posit that trusting beliefs have a direct effect on intention, leaving aside the afore-mentioned attitude to behavior, which used to be included in the original models of Fishbein and Ajzen (1980). Since Davis 1989, however, there has been an open discourse as to whether attitude has a place in this chain of effects (e.g. Li et al 2008) or whether it does not (e.g. Gefen et al. 2003, Wang and Benbasat 2005). In their general study of trust in new technologies, to which AI belongs, McKnight et al. (2002) leave out a trusting attitude and instead provide evidence that trusting beliefs have a statistically significant and direct effect on trusting intention. Extending these prior studies in our domain, we hypothesize that trusting beliefs also have a direct effect on trusting intention in our context of AI-based DSS.

H1: In an interaction with an AI-based DSS, trusting beliefs positively affect trusting intention.

As described in TRA, trusting intention leads to a certain behavior. Empirical testing of behavior is much more complex than, for example, a survey with questions about the participants' intentions. This is why empirical studies of behavior are rare, yet they are important to provide evidence for the link between trusting intention and behavior, especially since other empirical studies in areas like data privacy (e.g. Nordberg et al., 2007) or healthy living (e.g. Sniehotta et al., 2005) contradict this causal link. Yet whereas they speak of the "intention-behavior-gap", we contribute to this discourse by explicitly including behavior in our study. More specifically, we put the following hypotheses to an empirical test in the focus area of AI-based DSS.

H2: In an interaction with an AI-based DSS, trusting intention positively affects behavior.

AI-based DSS draw on expertise that is near enough unrivaled, so the user's domain-specific expertise is of considerable relevance. Research on trust shows that trust is deeper when the trustee's expertise (in our case, the competence of the AI system) surpasses the trustor's expertise (the competence of the user) (Doney et al., 1998). If the system's expertise is inferior, the user questions the system's ability, in which case the user would rather trust their own opinion. Therefore, it is particularly interesting to examine the interplay between the expertise of the user and the competence of the AI-based DSS. Doing so allows us to investigate if a user's trust in the AI system is lower when their own domain-specific expertise is higher. To facilitate a comprehensive examination of how expertise affects trust, we developed the following two hypotheses, H3a and H3b.

H3: In an interaction with an AI-based DSS, the user's domain-specific expertise negatively affects a) trusting beliefs and b) trusting intention.

XAI represents the extension of AI as it explains the latter's suggestions to the user, following strategies and methods from human-computer interaction research and the social sciences (Miller, 2019). XAI aims to enable the human user to trace the suggestions of an AI, the benefit of which is meant to be an increase in trust (Miller, 2019). To date, research has found that explanations indeed increase trust (e.g. Pu & Chen, 2006) and even foster the intention to act (e.g. Herlocker et al., 2000). In DSS research, though, this is an inconclusive issue. In the area of e-commerce, it seems to hold true, yet for clinical DSS, explanations do not increase trust (Bussone et al., 2015; W. Wang & Benbasat, 2008). In neither research area, however, is there any regard for real behavior since those studies represent dry runs. Rather than observe how the participants of their experiments behave, they make them read about a certain setting and then ask them to state what they would do. It is worth noting, though, that there can be a huge discrepancy between the intention to do something and the actual behavior (Sheeran & Webb, 2016). As a recent literature review in this topic area indicates (Miller et al., 2017), there is no scientific study that measures and observes real behavior in XAI. We have, therefore, made it a purpose of this study to investigate the influence of explanations not only on trust (H4a) but also on behavior (H4b).

H4: Explanations of an AI positively affect its user's a) trusting intention and b) behavior.

Explanations can differ greatly in execution as well as effect. One example is a study of clinical DSS (Bussone et al., 2015) which illustrates that explanations, depending on which side of an argument they represent, can lead to the self-reliance of the user or the over-reliance on the DSS. The social sciences have studied explanations in terms of structure, purpose, etc. for thousands of years (Miller, 2019). Aristotle (384 - 322 B.C.), for instance, emphasized the use of reasoning and argumentation (Stanford Encyclopedia of Philosophy, 2017). Of course, research in this field has come a long way since then, but to this day the effects of explanations are debated at great length (Thagard, 1989). Research like that of Miller (2019), which focuses on progress in the area of XAI, is highly relevant in that it links the social sciences with traditional IT-intensive research fields.

Explanations delivered by AI systems have three dimensions (Gregor & Benbasat, 1999). The first dimension is the presentation format in which an explanation can be text-based or multi-media, meaning it can be enhanced by graphics, images, or animations (Gregor & Benbasat, 1999). Obviously, the afore-mentioned XAI category 'visualization' uses a multi-media method. In contrast, approaches from the other two categories tend to use text in the shape of rules or natural language. Moving on to the second dimension of explanations delivered by AI systems, this brings us to the so-called provision mechanism which defines when and how an explanation is provided. It can be done either automatically, in which case an explanation is always provided, or when user-invoked, which means that it is only provided on the user's demand and said provision is fully under the user's control, or it can be provided intelligently, meaning that there is some kind of implemented logic within the AI system that determines whether or not to provide an explanation (Gregor & Benbasat, 1999). Yet it is the third dimension, the content type of an explanation in the area of XAI, that is most discussed. It favors tracing, justification, strategy, or terminology (Gregor & Benbasat, 1999). For a detailed explanation of these four types, please refer to Gregor and Benbasat (1999). Suffice it to say here that the current discourse mainly features tracing and justification (e.g. Schmidt & Biessmann, 2019). Tracing is the process of getting to the suggestion (Gregor & Benbasat, 1999). Justification is “only” concerned with the reason as to why a provided suggestion seems to be a good one (Biran & Cotton, 2017). If one is to understand an explanation by one of these two methods, tracing requires more background knowledge about AI in general, and indeed about the specific AI system in question. As for the practical application of AI, there are those (e.g. Biran & McKeown, 2017) who say that it is unreasonable to presume that real users like doctors, judges, bankers, or even average software developers have this kind of knowledge. Justification, on the other hand, offers reasons for a suggestion, rather than just a way of getting to the desired goal. Such reasoning can be adapted to the receiver of the explanation, which is in line with our usual human behavior of presenting explanations relative to the explainee’s beliefs (Miller, 2019). Automatically, we explain the same objective differently depending on the addressed person.

The same is true at the other end. Users also consume explanations differently depending on their domain expertise (Gregor & Benbasat, 1999). A novice self-improves in so far as they learn from an explanation, yet a novice can assess neither the suggestions of the AI nor their explanations properly (Gregor & Benbasat, 1999). Explanations cannot, therefore, contribute to the process of validating suggestions of the AI as far as a novice is concerned. Hence, giving a novice an explanation does not affect their acceptance of AI nor their trust in it. On the contrary, providing extra information by adding an explanation to the AI’s suggestion may even overstrain novices as there may be too much for them to process. For novices, then, explanations might even decrease trust. Experts, on the other hand, usually have prior knowledge and personal opinions, which is why they primarily use explanations to verify if suggestions match their opinions. Otherwise, they use them to resolve disagreements if their opinions and the AI’s suggestions are not in line (Gregor & Benbasat, 1999). As far as experts are concerned, then, we argue that explanations provide additional information that enriches the process of validating the AI’s suggestions and this greater use of the explanations along with the increased transparency of the AI fosters trust in it. Following this line of argument, we formulate hypothesis H5.

H5: In interactions with an AI-based DSS, the effect of an explanation on one's trusting intention is moderated by domain-specific expertise in the sense that explanations positively affect trust for people with high expertise and negatively affect trust for those with low expertise.

3.3.3 Experiment design and procedures

To test the hypotheses empirically, we perform an online experiment with a self-developed AI system for the game of chess. With the developed algorithm, we build two systems, one taking the role of the opponent, the other the role of the supporting AI. Both use the same underlying algorithm but different interfaces, and they perform separate roles for the user. The systems do not interact with each other directly, but merely indirectly in the sense that they both know the chessboard and the user's decisions. Due to wide applicability, controllability, and technical feasibility, we operationalize the explanations as automatic text-based justifications of the AI's suggestions.

We choose chess as the application domain of our experiment because AI and chess have a long and close relationship that has written numerous success stories (Silver et al., 2017). Our contribution to this field is an AI-based DSS that we created to support its users by suggesting chess moves. We have four reasons for doing so in this domain: The first reason is that a participant's behavior can easily be determined by said participant's chess move. The second reason is that, due to the uncomplicated data collection, this behavior is technically observable. The third reason is that, although there are various chess engines, such as Stockfish or AlphaZero (CEGT Team, 2020), there is no dominant winning strategy that can determine the best move in every situation. This uncertainty also applies to most real-world scenarios like deciding between two similar job applicants or medical treatments (cf. Holzinger et al., 2019). This brings us to the fourth and final reason. By virtue of the fact that anyone can become a skilled chess player, regardless of their educational background, this game represents an "excellent example of higher-order cognition" (van der Maas & Wagenmakers, 2005). As such, it is just as relevant when investigating domain-specific expertise for other scientific purposes like Human Problem Solving by Newell and Simon (Newell & Simon, 1972).

The structure of our experiment follows that of Li et al. (2008). Initially, we screen for participants with a minimum level of domain expertise (chess playing skill) and measure their level of expertise. Then, all participants make multiple moves in three different chess scenarios in which they play against a computer opponent while receiving helpful suggestions from the self-developed AI-based DSS. The level of explanation provided by the AI-based DSS is the between-subject treatment variable with random assignment to treatments. The measure of a participant's behavior is whether or not the participant decides to follow the AI's suggestions. The performance in these three chess scenarios is how we determine each participant's performance-based financial compensation. Finally, we conduct a subsequent survey to query perceptual variables, including concepts like trusting beliefs and trusting intention.

3.3.3.1 Recruiting and selecting participants

Our participants were recruited via mTurk, an online labor market frequently used for academic research [e.g., Freedman et al., 2020; Rahwan et al., 2019; Saxena et al., 2020). By providing easy access to a large, stable, and diverse subject pool, mTurk facilitated the selection we required for this research project (Mason & Suri, 2012). More specifically, since users of DSS typically have prior domain-specific knowledge (Ribeiro et al., 2016), we set a basic chess experience as a mandatory requirement for participation in our experiment. We ensured this in two ways, first by stating said requirement in the introductory text in which we asked for participation, and then by showing our potential participants two very easy boards and asking them to make the best move. If they failed in this task, we assumed that a basic understanding of chess rules could not be guaranteed and did not let them continue the experiment. The participants also answered questions about their demographics – age, degree of education, and type of occupation – but this information was only collected for descriptive purposes and not used to select participants.

3.3.3.2 Domain-specific expertise

The domain-specific expertise of our participants, meaning their level of expertise in chess, was assessed with the Amsterdam Chess Test (ACT) (van der Maas & Wagenmakers, 2005). The ACT measures chess skills in a high degree of detail by making test subjects perform five tasks: a choose-a-move task, a motivation questionnaire, a predict-a-move task, a verbal knowledge questionnaire, and a recall task. This thoroughness is required since this test is also used to validate top chess players. According to its authors (van der Maas & Wagenmakers, 2005), however, a good test of chess expertise can also be achieved by focusing on the choose-a-move task and making a subject perform several of those. For such a task, a certain game situation is presented in which the participant shall make the best possible move. For a proper evaluation, these situations are created in such a way that one move clearly surpasses all other possibilities (van der Maas & Wagenmakers, 2005). The ACT provides sets of such tasks, and we choose five to evaluate the chess expertise of the participants.

3.3.3.3 Behavior

For this study, we determine a participant's behavior based on their decision-making during the experiment. To measure behavior, participants play through three chess scenarios, each starting in the middle of a different game and with a different setup of chess pieces on the board. For each of the three scenarios, they play a sequence of five moves with the support of the AI-based DSS, while a computer opponent makes the five countering moves. This sequence of five moves is short enough to prevent the participants from becoming bored, yet long enough for them to become familiar with the scenario (Hafizoğlu & Sen, 2019). The total of 15 moves provides 15 observation points to study their behavior.

When it comes to making a decision, AI-based DSS may outperform a human in terms of concentration or calculative power and thus may help the user to make a better decision (Agrawal, 2018). A spell checker in text editing software, for instance, helps the user to write grammatically correct sentences, but the user still has the executive power to accept or reject the suggestion (i.e. the grammatical correction). This increases trust in the DSS (Mesbah et al., 2019). In this process, the user starts by manifesting their guess by typing the words, and the DSS assesses these. Nothing else happens in more advanced DSS, but those responsible in the decision-making scenario often do not manifest their guess even though they might have one. Nonetheless, there is a notable increase in user acceptance (Thagard, 1989) and perceived trustworthiness (McKnight et al., 1998) if the suggestion is consistent with prior beliefs. Since we aim to investigate how the justifying of the AI's suggestions impacts the user's trust, this poses a problem for us as uncontrolled accordance with the participant's prior beliefs could lead to a bias in our experiment. To prevent this, the participants start by making a chess move on their own. Afterward, the AI-based DSS acts in the form of a Supporting AI by suggesting a move that differs from the one by the participant, which forces the said participant to decide between the personal move and the suggested move by the Supporting AI.

In this experiment, every decision-making process has four phases, shown in Figure 3.3-2 and explained in the following.

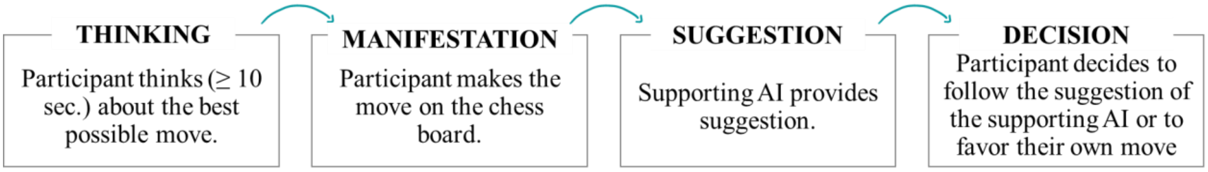

THINKING
Participant thinks (≥ 10 sec.) about the best possible move.
MANIFESTATION
Participant makes the move on the chess board.
SUGGESTION
Supporting AI provides suggestion.
DECISION
Participant decides to follow the suggestion of the supporting AI or to favor their own move

Figure 3.3-2: Phases in the decision-making process in the experiment

Each decision starts with the participant thinking about the setup on the board for at least ten seconds. By disallowing any move for ten seconds, we prevent participants from rushing into a decision. In the second phase, we make the participant act on their guess as to what may constitute the best move. Using drag and drop, the participant makes whatever move they please, as long as that move obeys the rules of chess.

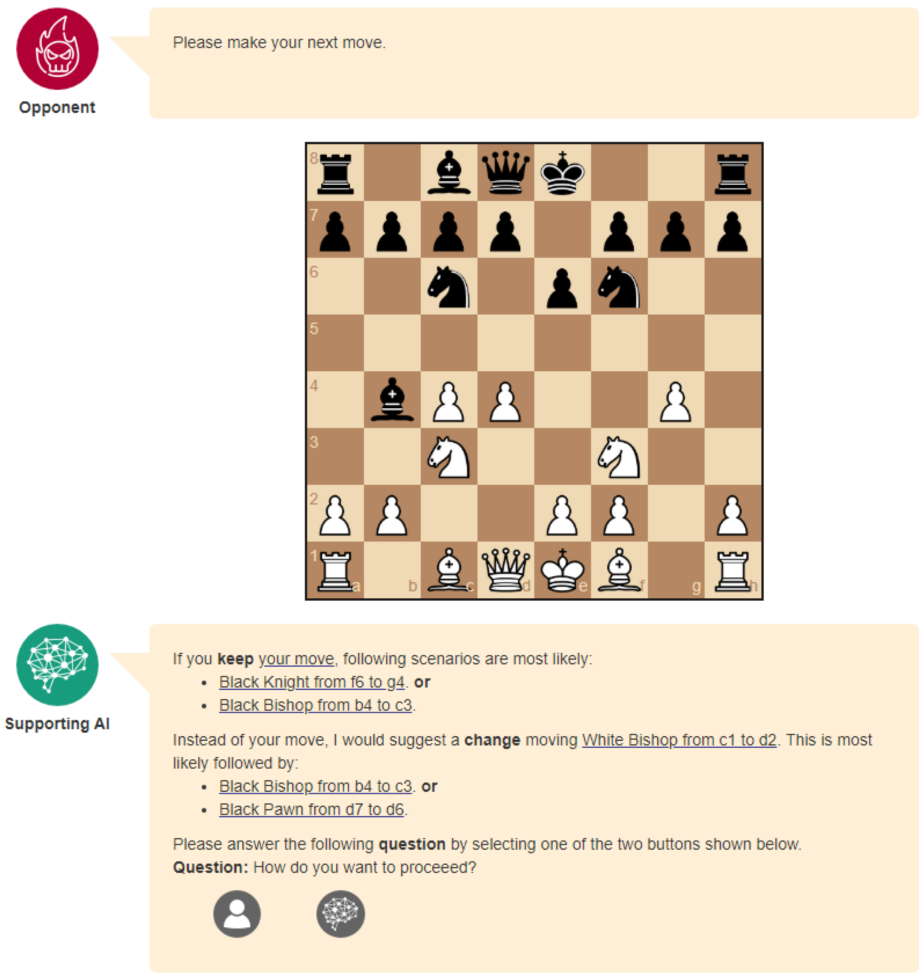
In the third phase, the supporting AI assesses the participant’s move and makes another suggestion. Thus, we create comparable experiences for all participants, as none of them get to agree with the supporting AI. However, this also means that the AI suggestion may be less competent. Imagine if the participant makes the move that the supporting AI assesses to be the best one. Since the supporting AI has to suggest an alternative move, it can only suggest the second-best move. This can negatively impact the perceived competence as well as the perceived trustworthiness of the supporting AI, which is why we consider it separately in the later data analysis. Most of the time, however, the AI-based DSS suggests a move that is superior to the participant’s personal decision, details of which are provided in the Results section. The third phase also includes the different treatments, which is to say whether or not the suggestion of the supporting AI is explained As the participant is asked to state their trust in the supporting AI, it might seem confusing if the supporting AI explains some suggested moves and others not. To prevent this, we favor a between-subject design that distinguishes between ‘no explanation’ and ‘explanation’. As noted, though, the design specifics for good explanations are not entirely clear. With this in mind, weincrease the robustness of our study by providing two different explanations, referring to them as A and B. They differ by the number of causes within the explanation, which is in accordance with the experiment of Lombrozo (2007). However, as the domain of this experiment is chess, rather than an alien disease, we define the number of causes as future moves (Lombrozo, 2007). This is consistent with commercial chess support systems such as Fritz of ChessBase as they offer guidance by stating the next most probable moves (ChessBase, 2020). In our experiment, explanation A is given with reference to the next move, whereas explanation B is given with reference to the next two moves. This is done accordingly in each case. Figure 3.3-3 illustrates an example of a chessboard and the way in which the supporting AI provides explanation A for its suggestion. Explanation B would state as follows:

If you keep your move, the following scenarios are most likely:

- Black Knight from f6 to g4, which can be answered by White Rook from h1 to g1 or
- Black Bishop from b4 to c3, which can be answered by White Pawn from b2 to c3

Instead of your move, I would suggest a change, moving White Bishop from c1 to d2. This is most likely followed by:

- Black Bishop from b4 to c3, which can be answered by White Bishop from d2 to c3
- Black Pawn from d7 to d6 which can be answered by White Pawn from d4 to d5



Please make your next move.

If you **keep** your move, following scenarios are most likely:

- Black Knight from f6 to g4. **or**
- Black Bishop from b4 to c3.

Instead of your move, I would suggest a **change** moving White Bishop from c1 to d2. This is most likely followed by:

- Black Bishop from b4 to c3. **or**
- Black Pawn from d7 to d6.

Please answer the following **question** by selecting one of the two buttons shown below.
Question: How do you want to proceeed?

Figure 3.3-3: Screenshot of the game scenario exemplifying explanation A of a suggestion

The fourth phase represents the participant’s decision and thus their behavior as they state the choice by clicking on one of those two buttons. Afterward, the opponent playing black makes its move, and the decision-making process starts all over.

Ultimately, there are 15 decisions to make. We aggregate these in a single variable that summarizes behavior by calculating the share of the decisions for which the participant followed the suggestions of the AI-based DSS. This share ranges from zero to unity. A value of 20% indicates that the participant chooses to make the move of the supporting AI rather than their own move three out of 15 times.

For this experiment, we developed two systems. In one, the supporting AI serves as DSS for the participant, in the other, an AI serves as the opponent. When optimizing the supporting AI, we followed the design of commercial chess support systems such as Fritz of ChessBase (ChessBase, 2020). These systems provide guidance by stating the next most probable moves. To mirror this procedure, we chose the tree-based algorithm negamax. This is a variant form of minimax search, optimized for a two-player game with an established simple evaluation function (Althöfer, 1990; CPW Team, 2018). This method allowed us to reverse-engineer the search tree and infer the next possible moves. As for the technical infrastructure, here we relied on an Amazon Web Service Lambda function and calculated the moves live during the experiment. Due to technical restrictions, however, we set the depth of the tree to 4. Finally, we designed the opponent system according to the same logic yet did not infer the next possible moves.

3.3.3.4 Post-Survey

Once both their expertise and their behavior were apparent from their decisions as to whether or not they followed the advice of the AI-based chess-play DSS, participants answered a survey on the remaining constructs of our research model. We used validated survey scales from prior research for every construct, including the scales of Li et al. (2008) to test for 'trusting intention' and the second-order construct 'trusting beliefs' with the first-order constructs 'benevolence', 'competence', and 'integrity'. Gefen et al. (2008) state that new users rely more on trust, "whereas more experienced users rely more on perceived usefulness" (p. 277) when making decisions about a system's use. To implement the necessary controls, we followed in the footsteps of previous researchers (e.g. Gefen et al. 2003) and included the standard technology acceptance constructs 'perceived ease of use' and 'perceived usefulness' as we used the scales by Davis (1989). For all items, we used 7-point Likert scales, ranging from strongly disagree (= 1) to strongly agree (= 7). When we conducted our survey, we ran through the constructs in the reverse order to their chain of effects. The order of the items was randomized. Please refer to Appendix A for those survey items.

Participants were compensated for their time, which on average was 21.7 minutes, by a fixed payment of 100 US Cent and a performance-based payment up to 60 US Cent. The average total payment was 120.7 US Cent, which is in the range of standard compensations on mTurk.

It is worth noting that participants of such studies may try to maximize their hourly earnings by rushing through the experiment without properly engaging with the instructions and reflecting on their answers. To reduce such noise in the data, we included two check questions, asking all of our participants to mark “disagree” or “the most right box”. The data of participants who did not correctly answer these questions was dropped.

3.3.4 Results

3.3.4.1 Data description and descriptive analysis

One hundred individuals participated in our experiment once at least basic chess knowledge was ascertained and they had all passed the attention checks. Since domain-level expertise is a key element of our research model, we split the participants into two groups, depending on their performance in the test of their domain-specific expertise. Group low (n=54) includes participants who made one out of five correct moves in the test, whereas group high (n=46) includes those who made at least two. So as to increase the robustness of our study, we had two types of explanations: A (one cause) and B (two causes). We could not find any significant difference in the averages between those treatments in regard to trusting beliefs (t-test, p = .612), trusting intention (t-test, p = .983), behavior (t-test, p = .761), and 2nd-rate (t-test, p = .228). Therefore, we do not further distinguish between the explanations A and B.

We combine the availability of an explanation or the lack thereof with the two levels of expertise and divide the participants into four groups. The groups are identified by a two-letter label, **LN** for **l**ow expertise and **n**o explanation. Table 3.3-1 shows the number of participants and demographic information of each group. The groups are homogenous with respect to the participants’ age (Kruskal-Wallis-Test, p = .198), gender (χ^2 test, p = .302), and the split between academics (participants who have a least a bachelor’s degree) and non-academics (χ2 test, p = .316).

Group	Expertise	Explanation	Number of Participants	Mean age	Gender in %			Academics in %
					m	f	d	
LN	Low	No	21	36.52	52	48	-	86
LE	Low	Yes	33	34.00	73	27	-	76
HN	High	No	11	34.73	55	45	-	91
HE	High	Yes	35	36.63	71	26	3	69
Overall			100	36.07	65	34	1	77

Table 3.3-1: Descriptive statistics of the participants

Table 3.3-2 illustrates the collected data per group. It includes 'trusting beliefs' and their dimensions 'integrity', 'benevolence', and 'competence' as well as their merged data, titled *as one*.

Group	Trusting Beliefs				Trusting Intention	Behavior in %	2^{nd}-rate in %
	Integrity	Benevolence	Competence	as one			
LN	5.58	5.43	5.67	5.56	5.63	46	16
LE	5.55	5.48	5.89	5.64	5.31	35	16
HN	5.16	4.95	5.12	5.08	4.14	47	22
HE	5.21	4.99	5.19	5.13	4.87	38	26
Overall	5.39	5.24	5.51	5.38	5.10	40	20
Scale	1 (low) to 7 (high)					0 (low) to 100 (high)	
Source	4 Items	2 Items	3 Items	composite	4 Items	Direct Measure	Direct Measure

Table 3.3-2: Descriptive statistics of the constructs[16]

When we say *behavior* in this context, we mean the behavior of going along with the alternative move suggested by the supporting AI. We measure this in percentage form, calculated from the number of times a participant does this in the series of fifteen moves. Accordingly, a value of 40% means that the participant favored the suggestion of the AI over their own move in six out of fifteen situations. As the suggestion of the supporting AI always differs from the user's opinion, *2^{nd}-rate* represents the share of moves in which those two would have agreed. It also represents the percentage of second-best suggestions. The overall average of 20% signifies that the supporting AI suggested three of fifteen moves that would have been the same as that of the participant. Conversely, it also signifies that in 80% of the cases, the AI identified a move seemingly superior to the move of the human player. The AI suggests the second-best move

[16] Due to issues with the discriminant validity, we dropped two items (c.f. section 4.2). Table 3.3-2 illustrates the subsequent data.

when the user has already made the best move, according to the AI. This means that a higher 2nd-rate indicates that the supporting AI was, on average, less competent. In other words, the higher the 2nd-rate, the more often the user made the move that the AI would have suggested as the best possible option. It comes as no surprise, then, that the 2nd-rate has a deeply negative association with the AI's perceived competence (Pearson correlation test, -.328 with p < .001). Further to be expected is the fact that the participant's expertise has a notable positive impact on the 2nd-rate (Pearson correlation test, .340 with p < .001), which shows that the supporting AI has more in common with a chess expert, rather than with a chess novice.

To understand potential biases in the interaction with the self-developed AI-based DSS and to make our analysis more robust, we asked participants to report their perceived ease of use along with their perceived usefulness of the system. No significant differences were noted – neither in perceived ease-of-use (t-test, p = .651) nor in perceived usefulness (t-test, p = .469) – between participants who received an explanation and those who did not. The same was true when we investigated the different levels of expertise individually as low (.856 and .522) and high (.513 and .975).

3.3.4.2 Structural equation modeling

For our model, we conceptualized 'trusting intention' as a first-order construct with a reflective multi-item measurement model. We modeled 'trusting beliefs' as a second-order construct with a formative measurement model and composite indicators to account for the three dimensions 'integrity', 'benevolence', and 'competence' (McKnight et al., 2002). We also treated these three dimensions as first-order constructs with reflective multi-item measurement models. 'Behavior', 'explanation', and 'expertise' are manifest variables.

For analyzing our model, we decided to use variance-based structural equation modeling (SEM), also referred to as partial least squares SEM (PLS-SEM), rather than covariance-based SEM (CB-SEM). We made this decision for two main reasons: First, we chose PLS-SEM because 'trusting beliefs' requires a formative measurement model with composite indicators. In principle, it is also possible to create such a model with CB-SEM (Bollen & Diamantopoulos, 2017; Grace & Bollen, 2008), but due to identification issues, this approach imposes severe limits on the interpretability of the resultant structural model (J. F. Hair et al., 2018). PLS-SEM, on the other hand, allows adequate modeling of the composite indicators (J. F. Hair et al., 2018) while simultaneously accommodating the reflective measurement models of other constructs (Roldán & Sánchez-Franco, 2012). PLS-SEM results in "practically no bias" when estimating the data of a model population, regardless of whether the measurement models are reflective or formative (Sarstedt et al., 2016). However, this only applies for PLS-SEM and not for its modified version, the consistent PLS-SEM (Dijkstra & Henseler, 2015; Sarstedt et al., 2016). With this in mind, we agree with Hair et al. (2018, p. 21) that PLS-SEM is “optimal for estimating models while simultaneously allowing approximating common factor models with effects indicators with practically no limitation.” The second reason to favor PLS-SEM over CB-SEM concerns data set considerations, as we examine 100 cases in our experiment. PLS-SEM makes lower demands on sample sizes than CB-SEM (J. F. Hair et al., 2019; Urbach et al., 2010). Having summarized prior studies, Urbach and Ahlemann (2010) suggest a rule of thumb for PLS-SEM according to which the minimal recommendations for sample size are 30 to 100 cases. In contrast, for CB-SEM, minimal recommendations range from 200 to 800 cases (Urbach et al., 2010).

To select further modeling specifics, we used the two-stage approach for moderation, seeing as a recent study shows that this approach outperforms alternatives, such as the orthogonalizing or the product-indicator approach (Becker et al., 2018). Furthermore, there are several ways to handle a higher-order construct in PLS-SEM, yet the (extended) repeated indicators and the two-stage approach are most prominent (Ringle et al., 2012; Sarstedt et al., 2019). For reflective-formative higher-order constructs, the former entails fewer biases in the estimation of the measurement model for higher-order constructs and the latter results in better parameter recovery of paths (Becker et al., 2012). Aiming for a better understanding of the connections between the constructs, we chose a two-stage approach. Given that its two variants, the disjoint and the embedded two-stage approach, achieve similar results (Sarstedt et al., 2019), there is no objective reason to prefer one, which is why we used the embedded two-stage approach due to personal preference. We used the software SmartPLS 3 (Ringle et al., 2015) to estimate path coefficients, calculate metrics, and adhere to the rules of Hair et al. (2017)as we chose 5,000 subsamples.

In line with the advice of Hair et al. (2019) on how to assess reflective measurement models for PLS-SEM, Table 3 displays the metrics based on the first stage of the embedded two-stage approach. This also includes a bootstrapping with 5,000 subsamples and the lower (5%) and upper (95%) bounds of the confidence interval. As for discriminant validity, recent studies have provided evidence that the traditional Fornell-Larcker criterion is less suitable than the heterotrait-monotrait (HTMT) method (Fornell & Larcker, 1981; Franke & Sarstedt, 2019; Henseler et al., 2015). The HTMT values should be below a threshold of .90 and significantly smaller than 1. As this was not the case, we addressed this issue with a common procedure and dropped one item for integrity and one for competence (c.f. Appendix A) (Henseler et al., 2015). Table 3 illustrates the subsequent values. For all constructs, the loadings (c.f. Appendix A) are above .708, ensuring the item reliability. For the internal reliability, ρA proposed by Dijkstra and Henseler (2015) is not only a more exact measure than traditional metrics, such as Cronbach's alpha (J. F. Hair et al., 2019), but it also makes it possible to estimate the confidence intervals via bootstrapping. Since all mean values for ρA are between .70 and .95, we can ascertain good internal reliability. In terms of convergent validity, Table 3 shows that the values for the average variance extracted (AVE) are above the proposed threshold of .50 (J. F. Hair et al., 2019).

		Integrity (I)			Benevolence (B)			Competence (C)			Trusting Intention (T)		
		5%	mean	95%	5%	mean	95%	5%	mean	95%	5%	mean	95%
ρ_A		.837	.883	.922	.685	.771	.843	.845	.898	.938	.837	.883	.920
AVE		.666	.739	-	.743	.804	-	.757	.829	-	.653	.725	-
HTMT	I	-	-	-	-	.889	.998	-	.833	.919	-	.677	.828
	B	-	.889	.998	-	-	-	-	.899	.975	-	.626	.806
	C	-	.833	.919	-	.899	.975	-	-	-	-	.752	.864
	TI	-	.677	.828	-	.626	.806	-	.752	.864	-	-	-

Table 3.3-3: Assessment of the reflective measurement models

In the second stage of the embedded two-stage approach, we validate the reflective-formative second-order 'trusting beliefs' in accordance with Sarstedt et al. (2019). A special focus lies on the lower order constructs (integrity, benevolence, competence) which we have already addressed in the previous paragraph. All of them are significant (p < .001) indicators with rather low weights (integrity: .357; benevolence: .318; competence: .425). Indicators with a significant but low weight of .5 and below can be deleted unless there are strong theoretical reasons that might affect content validity (J. F. Hair et al., 2019). As all of our indicators have weights below .5 and concern scientifically established dimensions of trusting beliefs (Gefen et al., 2003; Li et al., 2008; McKnight et al., 2002; W. Wang & Benbasat, 2005), we decided to keep all of those indicators. For convergent validity, we calculated the AVE as .824 with a lower 5% bound of the confidence interval of .759, which exceeds the threshold of .50 (J. F. Hair et al., 2019). To assess collinearity between the indicators, we calculated the variance inflation factor (integrity: 2.686; benevolence: 2.719; competence: 2.824), which is "ideal" because the values are below the conservative threshold of 3 (J. F. Hair et al., 2019).

Finally, Figure 3.3-4 illustrates the resulting SEM (for further details please refer to Appendix B). As hypothesized, it shows that 'trusting beliefs' has an significant positive effect on trusting intention (H1), much in the way that trusting intention has a significant positive effect on behavior (H2), whereas the user's expertise has significant negative effects on trusting beliefs (H3a). The direct effect of the user's expertise on trusting intention (H3b) does not differ significantly from zero, but the total effect does (-.284 with p < .01). The statistical model does not support our theoretical hypotheses concerning the direct positive effects that an AI system's explanation has on either trusting intention (H4a) or behavior (H4b). Finally, we hypothesized that the user's expertise moderates the effect that explaining the suggestion of an AI-based DSS has on trusting intention. The data supports this hypothesis H5: for users with higher expertise, receiving an explanation does increase the trusting intention. This is particularly interesting as expertise on its own negatively affects trusting intention. This negative direct effect of expertise on trusting intention can, at least in part, be offset by explaining the suggestion of the AI-based DSS.

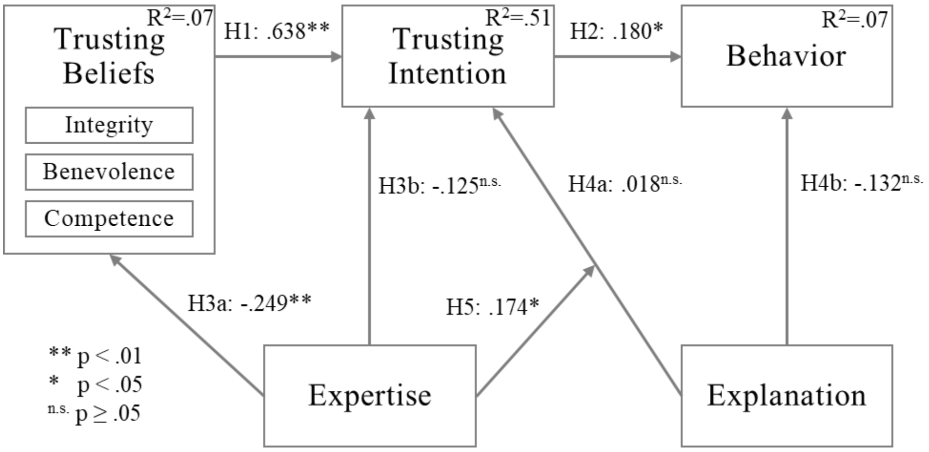


Figure 3.3-4: Estimated model

With regard to the explanatory power, Figure 3.3-4 illustrates the R2 values reporting the in-sample prediction (Hair Jr, 2020). The value of .07 for behavior indicates that the model explains 7% of the variability of behavior. For predictive power, we refer to PLSpredict (Shmueli et al., 2016) as it uses multiple holdout samples selected at random (Shmueli et al., 2019). As such, it facilitates "out-of-sample prediction for a single theoretical model in addition to in-sample prediction" (Hair Jr, 2020, p. 2). In this process, we obey the rules of Shmueli et al. (2019), choosing 10 as the number of folds and 10 as the number of repetitions. As a result, the value of Q2predict representing a naïve benchmark is slightly negative at -.02. This indicates that the predictive relevance is not confirmed (Shmueli et al., 2019).

3.3.5 Discussion

Our paper contributes to theory as well as to practice with three main implications which we outline in the following. Before we do so, however, we would like to outline an interesting observation about our data and point out a design dilemma.

With regard to the observed behavior, it is remarkable that the average rate for following the AI was about 40%, which means that, on average, users only took the advice of the supporting AI on 6 of 15 occasions. Meanwhile, we determined an average of 20% for the 2nd-rate, i.e., the rate at which the AI proposed not the best, but rather the second-best move. This means that 80% of the suggestions, and thus 12 of 15, are truly recommendable. These numbers result in a surprisingly small maximum following ratio of 6 of 12. This behavior is surprising, especially in the game of chess, as computers can by far surpass humans in this arena and this has been a widely known fact ever since the famous victory of Deep Blue against the reigning world champion in 1997 (Washington Examiner, 1997). Besides, this happened before AI received much greater exposure in the media landscape due to widely publicized technical advances, such as AlphaZero (Silver et al., 2017). Given these developments, we expected a much higher rate of users making the moves suggested by the AI. Perhaps participants did not expect our AI to perform as well as Deep Blue or AlphaZero, so it might be understandable that they were more reluctant to follow the suggestions of the AI in our experiment. However, our participants attributed high competence to the supporting AI, the average being a 5.5 on the 7-point Likert scale, which indicates that underestimating the AI is not a valid explanation as to why they rarely followed its advice, but as with human behavior in general, it is difficult to predict since there are seemingly countless influencing factors. In terms of predictive power, then, our model falls somewhat short.

We also faced a design dilemma in our experiment. Since human users are prone to cognitive biases that affect their decisions (Acciarini et al., 2020; Bowes et al., 2020), the manner in which the AI-based DSS makes its suggestion and the point at which it does so in the decision-making process are of notable relevance. For an experimental study, such biases are obstacles that have to be minimized or controlled. With regard to decision-making, there are two design possibilities. The AI decision support can be given either before or after the user assesses the situation and forms an opinion.

If this is done ‘before’, multiple biases can affect the decisions, such as the default effect. This is the tendency to favor the default option (Anaraky et al., 2020) and thus the proposed suggestion of the DSS. Another potential bias is the hindsight bias. This is the technical term for the human tendency to perceive an event “more predictable after it becomes known than it was before it became known” (Roese & Vohs, 2012, p. 411). Accordingly, the human user is likely to approve the suggestion of the AI-enabled DSS even though it might contradict their own opinion. On the contrary, if the AI decision support is given ‘after’ and the system agrees with one’s opinion, one can feel acknowledged, which can feel like a compliment. This can foster a positive atmosphere which, in turn, can foster greater trust (McKnight et al., 1998). In that case, the user is more likely to accept a later suggestion of the AI even though it may conflict with their intuition. When the system disagrees with one’s opinion, however, it creates a notably different decision-making scenario that is affected by other biases, such as the escalation of commitment. Due to this bias, people stick to a choice they made despite understanding the logical implication that doing so might lead to undesirable consequences (Staw, 1996).

Regardless of the preferred design, then, cognitive biases affect decisions and cannot be eliminated. Therefore, we simulated these circumstances for the participants of our study by making them state their intuition about the best possible move first. Only after that did we let the AI system propose an alternative move. This might differ from real-world settings, as users do not always express their gut inner opinion before receiving advice from a DSS, but this way the same biases affect all participants. It is worth noting, however, that there is a risk associated with this method as it might intensify biases like the escalation of commitment. This may be the reason that explanations in our approach do not affect the user in general (H4a and H4b are rejected). Furthermore, if the user chooses the best move first, the system can only suggest the second-best move. As our results show, the percentage of second-best suggestions affects the perceived competence of the system. If a user has a high level of expertise and thus chooses the best move on their own accord, the quality of the AI-based DSS is lowered. In practice, however, a higher level of expertise automatically leads to a lower perceived quality of the suggestions of third parties as greater knowledge enables people to assess them correctly. One could argue, then, that the design of our experiment does not cause this effect but rather reinforces it (e.g., H3a). Nonetheless, the perceived competence has the highest average (5.5) among the trusting beliefs, which leads us to conclude that the second-best move is not a significant issue.

3.3.5.1 Contribution

With this study, we answer several calls for further research (e.g. Biran & Cotton, 2017; McKnight et al., 2002; Miller et al., 2017; Miller, 2019). We contribute to the literature on trust in AI systems as we examine previous research on trust in other technologies and adapt those approaches to AI-based DSS.

Closely related to the research on TRA and similar trust-related models, we provide empirical evidence for the relationship between trusting intention and behavior in a decision-making scenario. Discussions of these two concepts often raise questions about a supposed intention-behavior gap. We contribute to this discussion by measuring actual behavior, which has rarely been done in publications on XAI. We show that there is a link between intention and behavior, but the relationship is weaker than we expected. What our results also suggest is that more research ought to include measurement of behavior, rather than treat intention as a proxy for behavior.

Furthermore, we shed light on the interplay between expertise and explanations in the context of AI-based DSS. We show how this affects trust and trust-related behavior, which is crucial in expediting the spread of AI systems (Du & Xie, 2020).

In practical terms, our research helps companies working on AI solutions to better understand how users can build trust in their AI systems. For instance, software developers can take our insights about XAI as a starting point to build systems that adequately justify their decisions to users. Our results also show that it is important to closely study the target group of a system prior to its release, as it is conducive to its success if the provider assesses the need for explanations with respect to the users' level of expertise.

3.3.5.2 Implications of our results

1. implication: It is not enough to pick the low-hanging fruit – Investigating intention is only a mediocre proxy for studying user behavior.

Previous AI research has hypothesized antecedents of user behavior (Miller et al., 2017), for example, by examining trusting intention. However, not only is there plenty of evidence for the intention-behavior gap (e.g. Nordberg et al., 2007), but there is also a call for further research on the link between trusting intention and behavior (e.g. McKnight et al., 2002). We have answered this call with our study. We can now confirm the link between intention and behavior in the context of AI-based DSS, which is to say that our H2 is supported, but the effect is weak. This has far-reaching consequences for XAI research. It means that one can approximate the right way to affect user behavior via improved trust by examining the user's trusting intention, as most existing literature does, but this approximation is far from perfect. Indeed, trusting intention has a weak effect on behavior (H2: .180). Furthermore, the in-sample explanatory power for behavior is marginal (R2 = .07), and out-of-sample prediction is unreliable. Our research indicates that the considerably simpler design of user studies which only assess trusting intentions, rather than also observe user behavior, is an insufficient proxy. When the research remit includes user behavior, this behavior should be measured.

2. implication: Do not mess with experts – Expertise negatively affects trust in (explained) AI interaction.

Our results also show that the trusting belief and trusting intention of users decrease as domain-specific expertise increases, which is to say that our H3a is supported and a mediated effect can be noted with regard to our H3b. Everybody is likely to know the feeling when somebody makes a bad or indeed a wrong decision. However, the assessment of decision quality that creates this feeling is subjective, meaning that it depends on one's perspective. In the context of AI-based DSS, this means that users evaluate the suggestion of the system in relation to their prior beliefs and knowledge. So, if a user's knowledge or expertise improves, it becomes more difficult for the system to surpass the user in terms of expertise. Consequently, there is a rise in the percentage of perceived bad suggestions of the DSS.

This insight has far-reaching consequences for developers of AI systems. Those involved in the process of developing an AI-based DSS must not only have substantial technological skills but also domain-specific knowledge. Our findings suggest that experts have lesser trust than beginners, which means that developing AI-based DSS for experts is an even greater challenge. These days, numerous AI systems are developed for experts, for instance, to support physicians in making a diagnosis in the area of healthcare or to support judges in the remit of the judiciary. It is, then, of utmost importance to know the application area of an AI system to gain the users' trust. Developers must think beyond the usual, largely technological challenges and find ways of establishing the trust of the target group.

3. implication: Know your user – Explanation moderated by expertise affects trusting intention.

Our results show that explanation moderated by expertise affects trusting intention, which means that our H5 is supported. Again, we see the importance of a circumspect creation of AI systems that are intended to support experts. Users with high domain-specific knowledge are more willing to trust the system if an explanation is provided. In contrast, users with low domain-specific knowledge, who have less knowledge about the decision and thus probably struggle to truly understand an explanation, do not build more trust when an explanation is provided. Therefore, AI-system developers should not underestimate the importance of providing good explanations for decisions of AI-based DSS that are intended for users with high domain-specific knowledge. With regard to applications for non-experts, however, explanations can play a subordinate role in development.

3.3.6 Limitations, further research & conclusion

In this study, we investigated how expertise affects trust and behavior in human users of XAI systems. We analyzed previous research and developed five theoretical hypotheses. To evaluate those, we conducted an online experiment with a custom-developed AI-based DSS in the domain of chess. Our results show that the chain of effects – trusting beliefs, trusting intention, and behavior – also applies to AI-based DSS. Additionally, we focused on the user and found support for our hypothesis that user expertise decreases the user's trust in an AI-based DSS. We also showed that explanations generally do not have a positive impact on trust or behavior in the context of XAI. When users have greater domain-specific expertise, however, explanations do matter.

At this point, it is worth observing that our study has certain limitations that other researchers would do well to explore. We conducted our experiment online, so its presentation was identical for each participant. However, we were unable to control the surroundings of those participants. Furthermore, our results are based on the data of 100 participants in the domain of chess. Given this relatively small sample size and the vagaries of gaming effects, our findings might be somewhat skewed. This may account for the insignificant effect that explanations had on trust (H4a) or behavior (H4b). Further worth noting is the fact that the perceived complexity of every single decision was not examined in this study. To minimize potential fallout from this methodological decision, we selected all the chess scenarios from the early middle game. However, perceived complexity can depend on familiarity with the opening which was not included and might, therefore, affect the results of this study, as might the validation of the second-order construct 'trusting beliefs', since our questionnaire does not include an alternative reflectively measure. Therefore, a redundancy analysis is not possible (Grace & Bollen, 2008).

Like those who went before us to research trust in technologies (e.g. Li et al., 2008; McKnight et al., 2002), we built our model on similarities between TRA and TPB. Upon introducing TRA (Fishbein & Ajzen, 1980), Ajzen (1985) developed the Theory of Planned Behavior (TPB) to improve predictive power beyond TRA. In the tradition of research into emerging technologies, we focused on the common factors of TRA and TPB and left aside their specifications. Future research may advance our model and investigate further variables of the more recent TPB, especially the role of perceived behavioral control. Another area of interest that those who come after us may wish to explore further is the explanation quality of the AI system. Recently, Holzinger et al. (2019) introduced a System Causability Scale for quality measurement at the interface between humans and AI. They proposed 10 suitable items for a potential survey, such as "I was able to use the explanations with my knowledge base" Holzinger et al., 2020, p. 196). In future research, this System Causability Scale could be used to assess the quality of AI explanations. Thus, it could be used to gain stand-alone insights or to extend our proposed model. By focusing on chess and rule-based systems, we deliberately chose a context for our study that is closely linked to AI in both theory and practice (Silver et al., 2017). Nevertheless, this context limits the extent to which general truths can be extrapolated from our specific results. As AI spreads ever further through modern life and different forms of AI system become available, such as rule-based systems and machine learning systems, it is certainly interesting to examine how our results can be transferred to other forms of AI (esp. deep learning) and indeed to other contexts like health care or the judiciary, since AI is now being used in both. A final point worth making here is that we focused our inquiry on expertise, yet users of an AI-based DSS are not exclusively characterized by their expertise but also by further user characteristics, such as personality traits and cultural background, all of which affect the impact that an explanation has in the context of XAI (Gefen et al., 2008). A good starting point for further research might be to create so-called personas. In the area of software development, this technique is used to image prototypes of a group of users that have specific characteristics, attitudes, and habits (Cooper, 1999).

To conclude, then, our findings indicate that the impact of explanations on trust and behavior in XAI is more complex than presumed. This is evidenced by our empirical results which show that explanations need to fit the user to affect trust. To overcome the human suspicion of AI, it is crucial to understand which measures and methods foster the trust of the respective type of AI user.

References

Acciarini, C., Brunetta, F., & Boccardelli, P. (2020). Cognitive biases and decision-making strategies in times of change: A systematic literature review. *Management Decision*. Advance online publication. https://doi.org/10.1108/MD-07-2019-1006

Ågerfalk, P. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems*, 29(1), 1–8. https://doi.org/10.1080/0960085X.2020.1721947

Agrawal, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press.

Ajzen, I. (1985). *From intentions to actions: A theory of planned behavior*. In J. Kuhl & J. Beckmann (Eds.), Action control (pp. 11–39). Springer. https://doi.org/10.1007/978-3-642-69746-3_2

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211.

Althöfer, I. (1990). An incremental negamax algorithm. *Artificial Intelligence*, 43(1), 57–65. https://doi.org/10.1016/0004-3702(90)90070-G

Anaraky, R. G., Knijnenburg, B. P., & Risius, M. (2020). Exacerbating mindless compliance: The danger of justifications during privacy decision making in the context of facebook applications. *AIS Transactions on Human-Computer Interaction*, 12(2), 70–95. https://doi.org/10.17705/1thci.00129

Becker, J.-M., Klein, K., & Wetzels, M. (2012). Hierarchical latent variable models in pls-sem: Guidelines for using reflective-formative type models. *Long Range Planning*, 45(5-6), 359–394. https://doi.org/10.1016/j.lrp.2012.10.001

Becker, J.-M., Ringle, C. M., & Sarstedt, M. (2018). Estimating moderating effects in pls-sem and plsc-sem: Interaction term generation*data treatment. *Journal of Applied Structural Equation Modeling*, 2(2), 1–21. https://doi.org/10.47263/JASEM.2(2)01

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *Xai workshop at the 26th international joint conference on artificial intelligence*, Melbourne, Australia.

Biran, O., & McKeown, K. (2017). Human-centric justification of machine learning predictions. In *26th international joint conference on artificial intelligence*, Melbourne, Australia.

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods,* 22(3), 581–596. https://doi.org/10.1037/met0000056

Bowes, S. M., Ammirati, R. J., Costello, T. H., Basterfield, C., & Lilienfeld, S. O. (2020). Cognitive biases, heuristics, and logical fallacies in clinical practice: A brief field guide for practicing clinicians and supervisors. *Professional Psychology: Research and Practice,* 51(5), 435–445. https://doi.org/10.1037/pro0000309

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. In *International conference on healthcare informatics,* Dallas, United States of America.

CEGT Team (Ed.). (2020). Cegt-ratinglist. http://www.cegt.net/40_4_Ratinglist/40_4_BestVersion/rangliste.html

ChessBase (Ed.). (2020). *Fritz 17 - the giant pc chess program, now with fat fritz*. https://shop.chessbase.com/en/products/fritz_17

Cooper, A. (1999). *The inmates are running the asylum*. Sams.

CPW Team (Ed.). (2018). *Simplified evaluation function*. https://www.chessprogramming.org/index.php?title=Simplified_Evaluation_Function&oldid=2101

Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. https://doi.org/10.2307/249008

Dietz, G., & Gillespie, N. (2011). Building and restoring organisational trust. *Institute of Business Ethics.* London.

Dietz, G., & Hartog, D. N. den (2006). Measuring trust inside organisations. *Personnel Review*, 35(5), 557–588. https://doi.org/10.1108/00483480610682299

Dijkstra, T., & Henseler, J. (2015). Consistent and asymptotically normal pls estimators for linear structural equations. *Computational Statistics & Data Analysis*, 81, 10–23. https://doi.org/10.1016/j.csda.2014.07.008

Doney, P., Cannon, J., & Mullen, M. (1998). Understanding the influence of national culture on the development of trust. *Academy of Management Review*, 23(3), 601–620. https://doi.org/10.5465/amr.1998.926629

Du, S., & Xie, C. (2020). Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities. *Journal of Business Research*. Advance online publication. https://doi.org/10.1016/j.jbusres.2020.08.024

Fishbein, M., & Ajzen, I. (1980). *Belief, attitude, intention and behaviour: An introduction to theory and research.* Addison-Wesley.

Fornell, C., & Larcker, D. (1981). Structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(3), 382–388. https://doi.org/10.2307/3151312

Fosso Wamba, S., Bawack, R. E., Guthrie, C., Queiroz, M. M., & Carillo, K. D. A. (2020). Are we preparing for a good ai society? A bibliometric review and research agenda. *Technological Forecasting and Social Change*, 120482. https://doi.org/10.1016/j.techfore.2020.120482

Franke, G., & Sarstedt, M. (2019). Heuristics versus statistics in discriminant validity testing: A comparison of four procedures. *Internet Research*, 29(3), 430–447. https://doi.org/10.1108/IntR-12-2017-0515

Freedman, R., Borg, J. S., Sinnott-Armstrong, W., Dickerson, J. P., & Conitzer, V. (2020). Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283, Article 103261. https://doi.org/10.1016/j.artint.2020.103261

Gefen, D., Benbasat, I., & Pavlou, P. (2008). A research agenda for trust in online environments. *Journal of Management Information Systems*, 24(4), 275–286. https://doi.org/10.2753/MIS0742-1222240411

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and tam in online shopping: An integrated model. *MIS Quarterly*, 27(1), 51–90. https://doi.org/10.2307/30036519

Grace, J. B., & Bollen, K. A. (2008). Representing general theoretical concepts in structural equation models: The role of composite variables. *Environmental and Ecological Statistics*, 15(2), 191–213. https://doi.org/10.1007/s10651-007-0047-7

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

Hafizoğlu, F. M., & Sen, S. (2019). Understanding the influences of past experience on trust in human-agent teamwork. *ACM Transactions on Internet Technology,* 19(4), Article 45. https://doi.org/10.1145/3324300

Hair, J., Hollingsworth, C., Randolph, A., & Chong, A. (2017). An updated and expanded assessment of pls-sem in information systems research. *Industrial Management & Data Systems,* 117(3), 442–458. https://doi.org/10.1108/IMDS-04-2016-0130

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

Hair, J. F., Sarstedt, M., Ringle, C. M., & Gudergan, S. (2018). *Advanced issues in partial least squares structural equation modeling*. Sage.

Hair Jr, J. F. (2020). *Next-generation prediction metrics for composite-based pls-sem*. Industrial Management & Data Systems.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust - the case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change,* 105, 105–120. https://doi.org/10.1016/j.techfore.2015.12.014

Henseler, J., Ringle, C., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science,* 43(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8

Herlocker, J., Konstan, J., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Computer supported cooperative work*, Philadelphia, United States of America.

Hilton, D. (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

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: The system causability scale (scs): Comparing human and machine explanations. *Künstliche Intelligenz,* 34(2), 193–198. https://doi.org/10.1007/s13218-020-00636-z

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainabilty of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery,* 9(4), Article e1312. https://doi.org/10.1002/widm.1312

Kim, G., Shin, B., & Lee, H. G. (2009). Understanding dynamics between initial trust and usage intentions of mobile banking. *Information Systems Journal*, 19(3), 283–311. https://doi.org/10.1111/j.1365-2575.2007.00269.x

Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., & Séroussi, B. (2019). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*, 94, 42–53. https://doi.org/10.1016/j.artmed.2019.01.001

Li, X., Hess, T., & Valacich, J. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems*, 17(1), 39–71. https://doi.org/10.1016/j.jsis.2008.01.001

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

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *31th conference on neural information processing systems*, Long Beach, United States of America.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's mechanical turk. *Behavior Research Methods*, 44(1), 1–23. https://doi.org/10.3758/s13428-011-0124-6

Mayer, R., Davis, J., & Schoorman, D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709–734. https://doi.org/10.2307/258792

McKnight, H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81

McKnight, H., Cummings, L., & Chervany, N. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review*, 23(3), 473–490. https://doi.org/10.2307/259290

Mesbah, N., Tauchert, C., Olt, C. M., & Buxmann, P. (2019). Promoting trust in ai-based expert systems. In *25th americas conference on information systems*, Cancún, Mexico.

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., & Sonnenberg, L. (2017). Explainable ai: Beware of inmates running the asylum. *ArXiv* Preprint ArXiv:1712.00547.

Mishra, J., & Morrissey, M. (1990). Trust in employee/employer relationships: A survey of west michigan managers. *Public Personnel Management,* 19(4), 443–486. https://doi.org/10.1177/009102609001900408

Newell, A., & Simon, H. (1972). *Human problem solving*. Pearson Education.

Nordberg, P., Horne, D [Daniel], & Horne, D [David] (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41(1), 100–126. https://doi.org/10.1111/j.1745-6606.2006.00070.x

NRZ (Ed.). (2017). Die angst vor neuer technik ist so alt wie die menschheit. https://www.nrz.de/wochenende/die-angst-vor-neuer-technik-ist-so-alt-wie-die-menschheit-id209190935.html

Pavlou, P. A., & Gefen, D. (2004). Building effective online marketplaces with institution-based trust. *Information Systems Research,* 15(1), 37–59. https://doi.org/10.1287/isre.1040.0015

Power, D. (2002). *Decision support systems: Concepts and resources for managers*. Quorum Books.

Pu, P., & Chen, L. (2006). Trust building with explanation interfaces. In *11th international conference on intelligent user interfaces,* Sydney, Australia.

Rahwan, Z., Yoeli, E., & Fasolo, B. (2019). Heterogeneity in banker culture and its influence on dishonesty. *Nature*, 575(7782), 345–349. https://doi.org/10.1038/s41586-019-1741-y

Rai, A., Constantinides, P., & and Sarker, S. (2018). Editor’s comments: Next-generation digital platforms: Toward human–ai hybrids. *MIS Quarterly,* 43(1), iii–x.

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

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?". In *22nd acm sigkdd international conference on knowledge discovery and data mining,* San Francisco, United States of America.

Ringle, Sarstedt, & Straub (2012). Editor's comments: A critical look at the use of pls-sem in "mis quarterly". *MIS Quarterly*, 36(1), iii–xiv. https://doi.org/10.2307/41410402

Ringle, C., Wende, S., & Becker, J.-M. (2015). *SmartPLS 3.* http://www.smartpls.com

Robnik-Sikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5), 589–600. https://doi.org/10.1109/TKDE.2007.190734

Roese, N., & Vohs, K. (2012). Hindsight bias. *Perspectives on Psychological Science*, 7(5), 411–426. https://doi.org/10.1177/1745691612454303

Roldán, J. L., & Sánchez-Franco, M. J. (2012). *Variance-based structural equation modeling*. In M. Mora, O. Gelman, A. L. Steenkamp, & M. Raisinghani (Eds.), Research methodologies, innovations and philosophies in software systems engineering and information systems (pp. 193–221). IGI Global. https://doi.org/10.4018/978-1-4666-0179-6.ch010

Rudin, C., Waltz, D., Anderson, R., Boulanger, A., Salleb-Aouissi, A., Chow, M., Dutta, H., Gross, P., Huang, B., Ierome, S., Isaac, D., Kressner, A., Passonneau, R., Radeva, A., & Wu, L. (2012). Machine learning for the new york city power grid. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 34(2), 328–345. https://doi.org/10.1109/TPAMI.2011.108

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (3rd ed.). Pearson Education.

SAE International (Ed.). (2018). *Sae international releases updated visual chart for its "levels of driving automation" standard for self-driving vehicles.* https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles

Sarstedt, M., Hair, J. F., Cheah, J.-H., Becker, J.-M., & Ringle, C. M. (2019). How to specify, estimate, and validate higher-order constructs in pls-sem. *Australasian Marketing Journal*, 27(3), 197–211. https://doi.org/10.1016/j.ausmj.2019.05.003

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with pls and cbsem: Where the bias lies! *Journal of Business Research*, 69(10), 3998–4010. https://doi.org/10.1016/j.jbusres.2016.06.007

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2020). How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence*, 283, Article 103238. https://doi.org/10.1016/j.artint.2020.103238

Schmidt, P., & Biessmann, F. (2019). Quantifying interpretability and trust in machine learning systems. *ArXiv* Preprint ArXiv:1901.08558.

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

Sheeran, P., & Webb, T. (2016). The intention-behavior gap. *Social and Personality Psychology Compass*, 10(9), 503–518. https://doi.org/10.1111/spc3.12265

Shmueli, G., Ray, S., Velasquez Estrada, J. M., & Chatla, S. B. (2016). The elephant in the room: Predictive performance of pls models. *Journal of Business Research*, 69(10), 4552–4564. https://doi.org/10.1016/j.jbusres.2016.03.049

Shmueli, G., Sarstedt, M., Hair, J. F., Cheah, J.-H., Ting, H., Vaithilingam, S., & Ringle, C. M. (2019). Predictive model assessment in pls-sem: Guidelines for using plspredict. *European Journal of Marketing*, 53(11), 2322–2347. https://doi.org/10.1108/EJM-02-2019-0189

Siau, K., & Wang, W [Weiyu] (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31(2), 47–53.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *ArXiv* Preprint ArXiv:1712.01815.

Sniehotta, F., Scholz, U., & Schwarzer, R. (2005). Bridging the intention–behaviour gap: Planning, self-efficacy, and action control in the adoption and maintenance of physical exercise. *Psychology & Health*, 20(2), 143–160. https://doi.org/10.1080/08870440512331317670

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Laulhé Shaelou, S., Patel, A., Ryan, M., & Wright, D. (2021). Artificial intelligence for human flourishing – beyond principles for machine learning. *Journal of Business Research*, 124, 374–388. https://doi.org/10.1016/j.jbusres.2020.11.030

Stanford Encyclopedia of Philosophy (Ed.). (2017). *Aristotle's logic*. https://plato.stanford.edu/entries/aristotle-logic/

Staw, B. (1996). *The escalation of commitment: An update and appraisal. In Zur Shapira (Ed.), Organizational decision making* (pp. 191–215). Cambridge University Press. https://doi.org/10.1017/CBO9780511584169.011

The Telegraph (Ed.). (2018). *Chinese businesswoman accused of jaywalking after ai camera spots her face on an advert.* https://www.telegraph.co.uk/technology/2018/11/25/chinese-businesswoman-accused-jaywalking-ai-camera-spots-face/

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12(3), 435–502. https://doi.org/10.1017/S0140525X00057046

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2020). The relationship between trust in ai and trustworthy machine learning technologies. In *Conference on fairness, accountability, and transparency*, Barcelona, Spain.

Urbach, N., Ahlemann, F., & others (2010). Structural equation modeling in information systems research using partial least squares. *Journal of Information Technology Theory and Application*, 11(2), 5–40.

van der Maas, H., & Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise. *American Journal of Psychology*, 118(1), 29–60.

Velloso, M. (2018). *Difference between machine learning and ai* [Tweet]. Twitter. https://twitter.com/matvelloso/status/1065778379612282885

Wang, P. (2008). What do you mean by “ai”? In *1st conference on artificial general intelligence*, Memphis, United States of America.

Wang, W [Weiquan], & Benbasat, I. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems*, 6(3), 72–101. https://doi.org/10.17705/1jais.00065

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

Washington Examiner (Ed.). (1997). *Be afraid*. https://www.washingtonexaminer.com/weekly-standard/be-afraid-9802

Yan, Z., Kantola, R., & Zhang, P. (2011). A research model for human-computer trust interaction. In *10th ieee international conference on trust, security and privacy in computing and communications*, Changsha, China.

3.4 Towards a systematic inclusion of ethical impacts in design and development of software: A framework for ethical software development

Abstract:

Research and practice are busy constructing software, in particular AI-software, with in part unpredictable but ethically significant behavior. There exists a lack of prescriptive knowledge for ethics by design that provides guidance for software development teams. Against this backdrop, we adopt design science research to develop a process model to consider the ethical impacts of software during software development. As justificatory knowledge, we build on an extensive literature review and analysis of data from semi-structured interviews. Our design artifact is an ethical software development process. For evaluation, we discuss its design specifications against its design objectives, compare it with competing artifacts and discuss its understandability, completeness, and real-world fidelity with industry experts. Our paper contributes to a theory of design and action, combining existing approaches and data from our interviews into an integrated process model to guide software development teams.

Keywords: ethical software development, artificial moral agent, artificial intelligence, AI, design science research

Authors: Sarah Bayer, Annika Fähnle, Henner Gimpel

Status: This article is a working paper.

3.4.1 Introduction

There is no doubt that software, especially with AI-components, can have moral impact. How can or should software engineers during the development processes deal with potential moral impacts of their software? AI software products are on the rise in multiple areas of our lives, for example, as personal voice assistants, smart thermostats, or decision support tools for the judicial system or policing. A critical point that is especially discussed with AI software products is their autonomous advancement and improvement (Russell et al. 2016). This self-evolution makes it more and more complex for users as well as for software engineers to foresee the outcomes of a system.

As of today, most researchers, especially in the domain of ethics, classify software as ethically oblivious – that is, software does not act as moral agent as we humans do (Allen et al. 2006; Fritz et al. 2020). Moral agents are beings whose behavior is "governed by moral standards", meaning they have moral obligations and are accountable for their actions (Fossa 2018; Himma 2009). The question arises whether software with ever increasing cognitive capabilities may become moral agents. With the constant rise of AI in software systems, there is an ongoing debate about the (future) possibility (can we build?) and desirability (should we build?) of so-called artificial moral agents (AMAs) (e.g. Himma 2009, Dignum et al.; Fossa 2018). AMAs are "AI systems able to incorporate moral reasoning in their deliberation and to explain their behavior in terms of moral concepts" (Dignum et al., p. 62).

Regardless of the question whether with the help of AI, software systems will ever be classified as moral agents on the same level as for instance adult humans, it is of utmost importance to pursue ethics by design, incorporating the question of how to program software systems "to behave acceptably" (Allen et al. 2006) or in other words "to guarantee that an agent's behavior remains within given moral bound" (Dignum et al., p. 60). This is relevant for software without intelligence, but gets even more important with a rising share of software inhibiting an AI component. A working group currently concerned with ethical concerns during system design is IEEE P7000. Sarah Spiekermann, co-chair of that IEEE working group, and a co-author recently published an article about value-based engineering for ethics by design (Spiekermann and Winkler 2020). Ethics by design concerns the "methods, algorithms and tools needed to endow autonomous agents with the capability to reason about ethical aspects of their decision" (Dignum et al., p. 61). Spiekermann and Winkler discuss challenges in ethical values in system design and offer useful requirements and recommendations.

It is to be expected that discussions about ethics in software in general will gain momentum with further development, increasing complexity, and spread of AI software that make it even more difficult for software developers to predict how the system will act in new situations and foresee potential outcomes (Allen et al. 2006). To prevent negative ethical consequences, software development teams need structural guidance to consider ethical impacts during the software development process. Otherwise, it is not excessive to fear that software will sooner or later cause harm to humans as it acts against ethical principles.

Our paper brings together two related, yet entirely distinct streams of research that contribute to this discussion. On the one hand, due to its inherent link to ethics, we consider research in the domain of AMAs, as a specific subcategory of software, and extract guidance for ethical software development processes. Although our research is not limited to AI software but embraces software development in general, we specifically include literature about AMAs as this research stream often includes ethical discussions. On the other hand, we analyze research about ethics in traditional software development processes. Nevertheless, as we show in detail in section 2, research in both areas remains scarce. To the best of our knowledge, there exists no holistic process model a software development team can use to explicitly consider potential ethical impacts of their software product during its development. Yet, such a model would improve software development in an increasingly important way. To fill this research gap, the aim of our research is to

> *develop an approach for ethical software development, named ethical software development process model (ESDP), that provides guidance on how software development teams should incorporate the software product's potential ethical impacts during the design and development process.*

To address this research objective, we adopt the design science research (DSR) paradigm (Gregor and Hevner 2013) and leverage both literature and expert interviews. Our artifact is developed for all kinds of software development processes, independent of an obvious ethical component at first glance and independent of an AI component. It is applicable for software development teams of any size and embeds all common development processes. It is not applicable for downstream evaluation of the ethicality of existing software and does not provide concrete advise for ethical decisions by the software but ensures that ethical implications are acknowledged during the development process.

The remainder of the paper is structured as follows: first, we provide the theoretical background for our research (section 2). Next, we explain the applied methodological approach of DSR (section 3), followed by a description of the design and development process of the artifact (section 4) and presentation of the artifact in section 5. We evaluate our ESDP in section 6, discuss our results in section 7 and conclude with section 8.

3.4.2 Theoretical background

We first provide the theoretical background on software development process modelling, followed by a short introduction to the relevant sub-areas of ethics (computer ethics and machine ethics with AMAs as a subcategory). Lastly, we provide an overview of the existing research-niche ethics in software development.

3.4.2.1 Software development process modeling

A software development process model is an abstract description of a software development process (Lonchamp 1993) which consists of activities, methods, practices, and transformations

ed at developing and maintaining software (Slaughter et al. 2006). The process is usually performed by a software development team, that does not necessarily consist exclusively of software developers but may also include members such as computer scientists, designers or non-technical product or project owners (Spiekermann and Winkler 2020). We summarize all members of the team under the term software engineering professionals.

A software development project is mostly structured along a software development life cycle (SDLC) and follows sequential development phases: First, in the requirements phase the software development team defines the requirements for the software. Based on these requirements, a first concept for the product is defined. After, the team develops the software code, implements it, and subsequently tests the software regarding its functionality and compliance with the requirements. If the testing leads to satisfactory results, the software is installed, and the development cycle concludes with its maintenance. Each phase consists of activities which may also be conducted in parallel (Krcmar 2015). The SDLC forms the basis for most software development process models which do not necessarily follow the development phases in sequential order but can be classified along two dimensions: the level of formalization and whether the development follows a sequential or iterative procedure. Popular strongly formalized, sequential software development process models are for instance waterfall-model, V-model, W-model, whereas strongly formalized, iterative models include the spiral model, prototyping, or the OO lifecycle-model. Extreme programming or SCRUM are examples for weakly formalized, iterative models (Krcmar 2015).

Based on the recognition that there are many different software development processes, we define the following design objective:

(DO.1) An approach to ethical software development should be applicable to any kind of software development process model.

3.4.2.2 Computer and machine ethics

In software development process modelling, ethics is inherently embodied – independent of the potential awareness or unawareness towards ethical aspects from software developers. Ethics address the various abstract concerns that arise when moral agents make reflective and responsible decisions about certain behaviors or actions, as for instance providing procedures for determining what actions are good or bad (Copp 2006; Moor 1985). Next to meta-ethics and normative ethics, applied ethics is a subcategory of ethics that focuses on concrete practical issues, such as abortion, animal rights, medical ethics, or computer ethics (Singer 1986; Copp, 2006; Moor 1985).

Computer ethics, as a stream of applied ethics, has developed from information ethics which originally focused on discussing issues of information or data confidentiality, reliability, quality, and usage (Himma and Tavani 2008). Computer ethics aim at analyzing the nature and social impact of computer technology (Moor 1985). It argues that traditional normative ethical theories do not provide sufficient guidance for answering the ethical questions that arise from the use of computer technology as it enables humans to act and behave in ways that were not possible before (Brey 2010; Johnson 2004; Moor 1985). Computer ethics cover a very broad field of research and until today, no consensus exists concerning its structure and main areas of concerns (see for example Floridi 2010; Himma and Tavani 2008; Johnson 2004; Mitcham 1995). Following Floridi (2010), computer ethics can be categorized into ethical issues in information society in general and ethical issues especially in artificial contexts.

Concerning the first category, the ubiquitous influence of information technology leads to specific ethical questions and problems with regard to social issues, such as ownership or intellectual property (Stahl 2010), rights issues, especially regarding rights to privacy and freedom of speech as well as their abuses (Johnson 2004; Sullins 2010), security issues like cybercrime (Arquilla 2010; Johnson 1985), and issues of professional conduct (Johnson 2004; Mitcham 1995). Computer ethicists dealing with issues of professional conduct aim at understanding the social responsibility of computer professionals, i.e., people employed and educated in development, maintenance, selling and use of computer technology. This social responsibility is subject to controversy as, from an occupational perspective, computer professionals often do not act purely self-dependently but as employees of organizations. Thus, they suffer from conflicts of interest and often may not be involved in decisions (Johnson 2004). The issue of questions and diffusion of responsibility again underlines the necessity of providing a model to software development teams that enables them to actively include ethical aspects into the development process and therefore take on responsibility.

Organizations try to face this challenge for example by introducing professional codes of ethics, which obligate computer professionals to practice their profession in a beneficial and respected manner. Well-known codes are the Professional Software Engineering Code of Ethics by the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers Computer Society (IEEE) (final version: Gotterbarn et al. 1997), the British Computer Society's Code of Conduct (British Computer Society 2020), or Microsoft's ethical principles (Microsoft 2020). The codes list important, universally valid responsibilities, but rather high-level and without direct reference to software development process modelling, e.g. "Design and implement systems that are robustly and usably secure" (Association for Computing Machinery 2018).

The second category of research in computer ethics deals with ethical questions and problems explicitly in artificial contexts, i.e., regarding artificially created computer technology artifacts and artificial environments. This includes issues regarding artificial intelligence, life and virtual realities, like questions about responsibility as well as applicability of societal norms (Allen 2010; Johnson 2004; Mitcham 1995) and questions about the ethics of information technology artifacts themselves which address issues of artificial moral agency (Wiegel 2010). This highly relevant topic is also taken up by organizations (e.g., Microsoft's AI principles) as well as supranational initiatives, as the OECD principles on AI and the European Commission's ethics guidelines for trustworthy AI shows. All of them suggest guidelines that are useful in their totality, but not operationalized to design and development of software (e.g. "AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being" (OECD 2020), "AI should be ethical – respecting ethical principles and values" (European Comission 2019), or "AI systems should treat all people fairly" (Microsoft 2020)) (André et al. 2019). Due to the bandwidth of ethical issues related to AI, an entirely new research area called machine ethics emerged.

Machine ethics extend computer ethics research towards the ethics and moral agency of machines themselves and seek to implement ethical decision-making capabilities into the machines' design (Allen et al. 2006). Prominent representatives of machine ethics include Allen et al. (2006), Anderson and Anderson (2007), Moor (2006), and Floridi and Sanders (2004). The research area faces a variety of challenges due to its inherently interdisciplinary character: For example, machine ethicists need to convince the software engineering professional community of the necessity of incorporating ethical principles into machines and find a common language to approach this joint research endeavor. Moreover, from a technological point of view, challenges concern whether ethical guidelines are computable, if yes, whether a single correct solution exists to an ethical dilemma and whether ethically acting machines, i.e. AMAs, are at all possible (Allen et al. 2006; Anderson and Anderson 2007; Floridi and Sanders 2004).

With our research on how to account for ethics during the software development process, we contribute to the field of applied ethics. Moreover, we aim to bridge the named gap between software engineering professionals and machine ethicists, not via answering questions like if AMAs are possible, but via bringing existing findings of machine ethicists together with software engineering professionals’ workflows.

3.4.2.3 Ethics in software development

In order to retrieve potentially relevant knowledge for our approach to ethical software development, we examine literature about ethics in software development. The body of knowledge on ethics in software development can be categorized into five research streams:

(1) development of design principles for ethical software development (e.g. Al-A’ali (2008), Collins and Miller (1994), Gotterbarn et al. (1997), Hameed (2009), and Cary et al. (2003));

(2) approaches on addressing the individual software engineer’s ethics (e.g. Brandenburg and Minge (2019), Jia and Xin (2018), Hameed et al. (2010), Génova et al. (2007), McNamara, Smith, Murphy-Hill et al. (2018), and Spinellis (2017));

(3) research about ethical issues during software development, how they can be detected and overcome (e.g. Judy (2009), Thomson and Schmoldt (2001), and Wallnau (2018));

(4) research about how ethics and morals can be integrated into the software development process, primarily from the viewpoint of software engineers (e.g. Aydemir and Dalpiaz (2018), Cary et al. (2003), Karim et al. (2017), Gotterbarn and Miller (2010), Lurie and Mark (2015), and Rashid, Moore, May-Chahal et al. (2015)).

(5) research about inherently ethical software, namely AMAs (e.g. (Allen et al. 2000; Gips 1995; Himma 2009; Moor 2006).

Of the identified literature from stream (1) - (4), only three papers propose concrete frameworks or process models of ethical software development (Cary et al., 2003; Aydemir and Dalpiaz, 2018; Lurie and Mark, 2015). Of those, Aydemir and Dalpiaz (2018) come closest to developing a software development framework that guides software engineers to consider ethical impacts of their product. However, they put their primary focus on how to ensure harmony between ethical values and conduct during software development as opposed to considering the software product's potential ethical impacts.

The fifth stream of research is concerned with facettes of AMAs. Among different types of software, AMAs are the ones that are most closely and obviously related to ethics. Moor (2006) distinguishes between implicit and explicit AMAs. Implicit AMAs act ethical because of their internal functions, i.e., they are implicitly constructed to promote ethical behavior or at least avoid unethical behavior. They have been programmed to possess virtues which dominate their behavior. Explicit AMAs, on the other hand, follow specific ethical principles of for example deontological or consequentialist nature that have been previously programmed to obey (Allen et al. 2000; Gips 1995). So-called full AMAs can make autonomous ethical judgements like adult humans and are capable to reasonably justify them and are therefore able to possess responsibility for an action or behavior. To possess full moral agency, an agent is commonly required to possess properties such as consciousness, free will and intentionality (Himma 2009; Moor 2006) and it is yet unclear whether artificial agents will ever fulfill these prerequisites and whether this is indeed desirable.

The existing body of knowledge concerned with the design and development of AMA can be categorized into five research streams: The first stream discusses whether the existence of AMAs is technologically possible (e.g. Dodig C. and Çürüklü 2012). Second, research addresses which normative ethical theories or other moral approaches are appropriate for implementing moral behavior in their design (e.g. Allen et al. 2000; Allen et al. 2005; Bello and Bringsjord 2013; Bogosian 2017; Wiltshire 2015). The third stream of literature is concerned with their computational implementability and proposes computational AMA models (e.g. Anderson et al. 2006; Cervantes et al. 2016; Honarvar and Ghasem-Aghaee 2009). The fourth literature stream proposes how to assess the behavior of an AMA ex-post (e.g. Allen et al. 2000). The last research stream (actually consisting of a single paper) explores how to design AMAs (Wiegel 2006).

Of the existing literature about AMAs, only one paper proposes any guidance on how to design (Wiegel 2006) by introducing design requirements and principles. However, this work explicitly disregards software development aspects ("[…] a set of design principles can be formulated (from which I omit the software engineering oriented ones)" (Wiegel 2006, p. 2)) and therefore does not provide guidance on how to technically develop AMAs.

Next to these five research streams about ethics in software development, the working group IEEE P7000 – Engineering Methodologies for Ethical Life-Cycle Concerns Working Group aims at establishing "a process model by which engineers and technologists can address ethical consideration throughout the various stages of system initiation, analysis and design." (IEEE P7000 Working Group 2020). So far, no final results are published, but the Vice Chair, Sarah Spiekermann, currently published an article with Till Winkler about ethics by design, that is among others build on the learning from the working group. The article suggests 16 recommendations (e.g. "Not only engineers, but also corporate leaders and a wide group of stakeholders need to be involved in value prioritization") and 12 requirements (e.g. "To envision values, the three grand ethical theories of the Western Canon for value elicitation must be used (Utilitarianism, Virtue Ethics and Duty Ethics) […]") for value-based engineering (VBE) as well as three guiding questions to envision values (e.g. "What are all thinkable positive and negative consequences you can envision from the system's use for direct and indirect stakeholders?") (Spiekermann and Winkler 2020). This article comes close to the aim of our paper, but intensively focus on values, without providing concrete guidance to software development teams ins terms of when to answer which question during the software development process. The framework clearly focuses on the consideration of values for software development but does not provide an integrated view of the whole development process.

In a similar direction, the research area of value sensitive design (VSD) is concerned with values in the design process. Among numerous definitions for values, a popular one is that values "refer to what a person or group of people consider important in life" (Friedman et al. 2013). VSD is one of the most popular approaches to account for human values in technology design (Winkler and Spiekermann 2018). In VSD, the aim of incorporating human values into the design process is achieved via three often iterative phases: a conceptual, empirical, and technical investigation (Davis and Nathan 2015; Friedman et al. 2013). The conceptual phase concerns for instance questions about affected stakeholders and prioritization of moral values. The second phase empirically enriches the search for answers from the conceptual phase, for instance via including analyses to evaluate the success of a particular design. The final phase focus on how technologies can or cannot be used to support human values with the aim of designing technology to support the values previously defined (Friedman et al. 2013; Manders-Huits 2011). Again, this research stream incorporates important thoughts on how to account for values in the design process but does not provide a holistic process model as guidance for software engineering professionals.

In sum, to the best of our knowledge, no holistic approach exists which guides software development teams to consider the ethical impacts of their software product during development. We therefore intend to bridge the identified research gap by designing a model that guides software development teams to consider the ethical impacts of software products during development.

Against the backdrop of the preceding section, we draw from the extensive literature reviews as justificatory knowledge (Gregor and Hevner 2013) and define the following design objective:

(DO.2) An approach to ethical software development should guide software development teams to consider ethical impacts of software products during development.

3.4.3 Methodology

Following the DSR reference process proposed by Peffers et al. (2007), our research includes the following steps: (1) Problem identification and motivation of the research problem; (2) Definition of the objectives for a solution; (3) Design and development; (4) Demonstration; (5) Evaluation; and (6) Communication.

Our research problem is identified and motivated at the beginning of this paper, deriving the research gap and explaining its relevance for research and practice. The model should be generic in order to ensure its applicability for development of software. Design and development of the artifact is conducted with data from two different approaches, a literature research and qualitative interviews. We conduct a literature research in order to examine the theoretical background for software development, computer and machine ethics, and ethics in software development. The results of this literature review serve as justificatory knowledge in order to derive design objectives (see section 1) and design principles (see section 3) for our artifact.

Furthermore, we conduct qualitative, semi-structured interviews with industry experts to derive further design principles (see section 3) as our approach should be used by software development teams in practice. This is sensible as knowledge communicated in interviews count as field knowledge, an acknowledged form of justificatory knowledge (Gregor and Jones 2007). We conduct 7 interviews, lasting between 30-40 minutes each, with people holding the profession of either software engineers, chief technology officers (CTOs), or software product owners, either self-employed or holding a position in an organization where they have overview of, responsibility for, and power of disposition concerning a software development project. See Appendix A for detailed information about our interviewees. We base the interviews on an interview guide to ensure a comprehensive coverage of the intended subject area (Rubin and Rubin 2011) and use a semi-structured approach to follow the flow of the conversation and remain open to new findings and unexpected turns (Myers and Newman 2007).

The interviews aim at three goals: First, assessing the extent to which the respective interviewee is already aware of the ethical impacts the products of their software development projects may have. Second, receiving insights into their current (or latest) software development project and how, if at all, they implement ethical considerations in the development process. Third, discussing how the experts would – in an ideal world – integrate ethical considerations into software development processes. We audiotaped each interview, transcribed it in standard verbatim, and analyzed the transcripts in a two-stage process following Miles, Huberman, and Saldana’s (2014) first and second cycle coding methodology with the help of the software MAXQDA. We choose this inductive approach as not enough previous knowledge exists to create an initial list of codes necessary and we were able to develop emerging categories during analysis (Corbin and Strauss 1990). In the first coding cycle, we assign in-vivo, process, and emotion codes (Miles et al. 2014), whereas we cluster these codes and assign them to pattern codes in the second coding cycle. Drawing from the established concept of theoretical saturation, we discontinue data collection after seven interviews when no new codes occur in the data (Glaser and Strauss, 1967; Urquhart, 2013).

After having finished the coding, we extracted relevant factors and developed design principles. In this, we again reviewed the literature summarized in the Theoretical Background section and related it to the factors arising from the interviews where possible. This theoretical integration strengthens our results and highlights that some aspects are novel while others have been identified and discussed before. As a final step of the design and development phase of our DSR approach, considering the aim and design objectives of our research, we used the extant knowledge from literature along with the knowledge derived from the interviews and the design principles to craft the ESDP which we present in section 4.

To evaluate our model, we follow Sonnenberg and vom Brocke’s (2012) framework of evaluation activities in DSR. The iterative framework integrates into the DSR process as each DSR activity is followed by an evaluation activity. In our research, we focus on the two ex-ante evaluations (Eval 1 and Eval 2) suggested by Sonnenberg and vom Brocke (2012), as those are applied before the artifact is constructed and are therefore appropriate as first evaluation for our model.

Eval 1’s goal is to confirm that the research question is a relevant DSR problem by demonstrating its novelty and importance for research and practice (Sonnenberg and vom Brocke 2012). Appropriate methods for applying Eval 1 are, among others, assertions, literature reviews or surveys. We address this evaluation activity in the extensive literature review underlying the theoretical background in section 1, where we analyze the existing body of research and highlight the research gap concerning a software development process model which guides software development teams in considering the ethical impacts of software. We argue that the research gap stimulates the need to extend prescriptive knowledge by designing an approach that guides software development teams. Eval 1 further requires to derive design objectives from justificatory knowledge to evaluate whether an artifact would resolve the research problem (Lehnert et al. 2016). We address this by building on the literature review presented in the theoretical background and data collected in semi-structured, qualitative interviews as justificatory knowledge for design and development of our artifact.

Taking an ex-ante perspective, Eval2 aims to validate the artifact’s design specifications from an artificial and naturalistic angle (Pries-Heje et al. 2008). For this, we first conduct an artificial evaluation by discussing our approach to ethical software development against its design objectives derived from justificatory knowledge. To evaluate whether the model contributes to the existing body of knowledge, we further discuss the characteristics of competing artifacts against the design objectives. As competing artifacts, we select the prescriptive design approaches described in section 1, which, to the best of our knowledge, include all existing approaches to designing artifacts for ethical software development.

In a succeeding step, we validate the artifacts design specifications from a naturalistic perspective by conducting two additional qualitative, semi-structured interviews with experts of different organizations. The aim of this evaluation activity is to assess how the industry experts regard the design specifications’ understandability, completeness, and real-world fidelity. To further validate the model’s real-world fidelity, the second evaluation interview simulates its implementation. This approach is reasonable as logical reasoning and simulations are recommended methods for applying the Eval 2 activity (Sonnenberg and vom Brocke 2012). We report the results of EVAL 2 in section 5.

3.4.4 Design and development process

As mentioned above, we took an inductive, two stage approach to coding the interviews. In the first coding cycle we singled out a quote, for instance *“A motive can be [...] individual characters who drive this forward, and a corporate culture” (Interviewee 7),* and in-vivo coded it *corporate culture*. Then, in the second coding cycle (after the other interviews have been coded as well), we assign related codes to patterns and clusters. In this particular case, 15 codes related to how the moral culture of the employing organisation motivates ethical behaviour. Hence, we named the cluster “Culture of moral responsibility” and assigned it to the category “Motivating factors”. This is how we proceeded with all 234 quotes arising in the first coding cycle and aggregated them to 21 codes in the second cycle.

Based on the analysis of the qualitative data collected in the interviews, we build four categories. The first two categories introduce factors that motivate and hinder ethical software development, respectively. The third category provides insights into procedural good practices arising from experiences with software development process models. The fourth category depicts experts’ recommendations concerning the ideal ethical software development process from research and practice. Description of each category is concluded by deriving design principles for our artifact. Note, that we develop the process model with the aim to guide the entire software development team through the development process.

3.4.4.1 Category 1: Motivating factors

Our data analysis of literature and interviews revealed six factors motivating ethical software development. The factors are explained in Table 15.

Factor	Description	Exemplary interview excerpts	Exemplary literature sources
Ethical vision statement	A clearly defined ethical (product) vision of what the product is supposed to achieve, what not, and why, including an ethical goal, intrinsically aligns the development process to fulfill the vision.	*"[…] it has been our goal from the very beginning that privacy, security or IT-related security is one of our highest values. […] And so, naturally, from the very beginning we designed the system in such a way that it would do justice to the goal". (Interviewee 3)* *"[…] it is about the definition, what is it actually what we do, what is the goal of the project? At this point, is this a goal that you can justify or is it morally questionable? […] what do we actually want to achieve? What are the client's motives? And are these motives relatable?" (Interviewee 7)*	*Aydemir and Dalpiaz (2018)*
Empowerment	The perception that their opinion is heard within their team and organization and that they have the power to influence the project's outcome motivates software engineering professionals to address their ethical concerns.	*"The engineers actually discuss a lot: 'Is that actually OK for the user, privacy perspective or ethical? […]' Normally, if there's even one thing in the feature that isn't so OK for the user, there are many engineers who would say, 'Hey, we can't do it this way'. And most of the time it goes up to the director and he makes the decisions." (Interviewee 1)*	-
Culture of moral responsibility	A strong organizational culture of moral responsibility (e.g. via organizational code of conducts, ethics committees, compliance rules) motivates employees to become aware of and take on responsibility for potential ethical impacts of their products.	*"A motive can be […] individual characters who drive this forward, and a corporate culture". (Interviewee 7)*	*Cary et al. (2003)* *Spiekermann and Winkler (2020)* *Thomson and Schmoldt (2001)*
Communication of responsibilities	A clear communication of responsibilities, inducing a feeling of personal responsibility for the software under development and driving considerations about how potential ethical impacts can be adequately met by design.	*„[…] when I do a project myself, it's my job to somehow deal with the implications that the thing has. […] From the idea to the implementation, everything is mine. But when I receive a project from clients, I have to think about it. Would I do that now? […] If not, then I just won't implement it." (Interviewee 2)*	*Gotterbarn et al. (1997)*
Long-term value creation	An organizational focus on long-term value creation, entailing the attitude that software products need to possess a positive ethical image to achieve sustainable customer acceptance. This includes the attitude that higher initial investment costs for ethical development pays out in the long run.	*"[…] because it reflects badly on us if someone is getting hurt because he used our software". (Interviewee 2)* *"[…] and if there is a negative connotation that comes up all the time, then it won't sell". (Interviewee 2)*	*Thomson and Schmoldt (2001)*
Leadership by example	Leaders who emphasize value of ethical software development and put a strong focus on potential ethical impacts motivate their employees to imitate their behavior.	*"If you have great people around you, they can tell you what to do. And leadership in [organization] is pretty good. […] they try to put the user first wherever possible. I think if you have the right leaders then you are also going in the right direction". (Interviewee 1)*	*Gotterbarn et al. (1997)*

Table 3.4-1: Motivating factors for ethical software development

Against the backdrop of the factors motivating ethical development, we define the following design principle (DP) for our approach to ethical software development:

(DP.3) An approach to ethical software development should incorporate the motivating factors of ethical software development.

3.4.4.2 Category 2: Hindering factors

We identified three hindering factors, as depicted in the following table.

Factor	Description	Exemplary interview excerpts	Exemplary literature sources
Lack of ethical awareness	A lack of general awareness for potential ethical impacts for software hinders ethical software development. Especially software products that are not directly associated with an end user seem abstract and their ethical impacts may not be apparent.	*"Nope, I've never considered it [software development] from an ethical point of view or maybe just unconsciously". (Interviewee 1)*	*Aydemir and Dalpiaz (2018) Cary et al. (2003) Gotterbarn et al. (1997) Lurie and Mark (2015 Wallnau (2018)*
Deficiencies in ethical education	An existing deficit of ethical topics in the university education of software engineers leads software engineering professionals to depend on themselves or the employing organization for ethical education.	*"Yes. In my studies I didn't do any moral or ethical stuff. […] And we also have training once a year at [organization] to be ethical, follow the code of conduct and stuff. It probably should start earlier than that because the companies choose how they set the code of conduct. Not the engineer". (Interviewee 1)*	*Aydemir and Dalpiaz (2018) Génova et al. (2007) Hameed (2009) Jia and Xin (2018) Judy (2009)*
Lack of overview of project	Individuals who work on fractional parts of a large project lack overview of the project's big picture hindering them from developing a sense of responsibility. Especially in large organizations, software development projects can be very complex and big and stretch across different departments in several countries.	*"(…) if I get a task as a developer and they say, 'take care of it', then it's just a relatively small task package. And to see a big ethical implication in that, you have to know a lot about the product." (Interviewee 2)*	-

Table 3.4-2: Hindering factors for ethical software development

Against the backdrop of factors hindering ethical software development, we define the following design principle:

(DP.4) An approach to ethical software development should be designed to overcome the factors hindering ethical software development.

Factors hindering and motivating ethical software development are not disjunct. Rather, they may influence each other in terms of hindering factors inhibiting motivating ones, or motivating factors overcoming hindering ones.

3.4.4.3 Category 3: Procedural aspects

The interviews and literature further revealed procedural good practices of ethical software development.

Factor	Description	Exemplary interview excerpts	Exemplary literature sources
Ethical considerations overarching all development process phases	Consideration of ethical issues during the entire development process, not in a single software development process phase.	*"Experience has shown that in the past, I have always reflected [on ethical issues] during it [the software development process]. Before, it's hard to see at the beginning, and I don't know what it's going to be like". (Interviewee 4)*	*Lurie and Mark (2015)*
Designation of an ethics expert	The designated expert for ethics should be permanently part of the development team and have an interdisciplinary education or coaching of ethics, technology, and management.	-	*Spiekermann and Winkler (2020)*
Taking the stakeholders' perspectives	Identification of major and minor direct and indirect stakeholders, including the societal perspective, e.g., via answering the question "Who uses the product in which context?". Followed by the analysis of how they will be affected or affect others. Includes identification of stakeholders' values. Stakeholders might even be included in (part of) the ethical development process.	-	*Cary et al. (2003)* *Collins and Miller (1994)* *Friedman et al. (2013)* *Gotterbarn and Miller (2010)* *Lurie and Mark (2015)* *Manders-Huits (2011)* *Rashid et al. (2015)* *Spiekermann and Winkler (2020)* *Wallnau (2018)*
Habitual software development process model	Software development teams value the flexibility and autonomy of following their habitual version of the software development process – often an iterative and agile approach – for ethical software development.	*"This is one of the upper maxims of Scrum that people go before processes. [...] I have a developer, he is the communicative type, who likes to jump into the PO [project owner] role and then becomes more conceptual. And I have one, it draws him down incredibly when he has to do something else besides coding. [...] So – in other words – we have to make sure that we keep people before processes and do it so that it works for everyone". (Interviewee 5)*	-
Simulations and constant reviews of outcomes	Implementation of constant reviews during the development process and simulation of possible scenarios, especially when developing self-reinforcing algorithms in order to analyze the impacts of different scenarios in different contexts of use.	*"[...] Here, we have simulated how it would be if we had made the other decision. How about we made the first decision and we had people who know about Counter-factual Reasoning and stuff like that, mathematicians with PhDs. It's not trivial and I think you can get these biases out of the way with that kind of approach". (Interviewee 6)*	*Cary et al. (2003)* *Gotterbarn et al. (1997)* *Rashid et al. (2015)* *Spiekermann and Winkler (2020)* *Wallnau (2018)*
Help and advice by compliance department, lawyers, and consultants	Next to the ethics experts, compliance departments, consultants, or lawyers should be ask for advice if necessary.	*"At that moment, I would seek assistance. We have the legal department, which is also the compliance department." (Interviewee 5)*	-
Monitoring and evaluation	Constant monitoring and evaluation of the system after release, e.g., to explore the context of use and potential ethical impacts.	-	*Davis et al. (1988)* *Krcmar (2015)* *Ruparelia (2010)* *Spiekermann and Winkler (2020)*

Table 3.4-3: Procedural aspects for ethical software development

Against this backdrop, we define the following design principle to account for current practices of ethical software development:

(DP.5) An approach to ethical software development should incorporate procedural good practices of ethical software development processes.

3.4.4.4 Category 4: Expert recommendations on ethical software development

Finally, research and practice revealed five recommendations concerning the ideal ethical software development process:

Factor	Description	Exemplary interview excerpts	Exemplary literature sources
Identification of potential ethical impacts of software product	Identification, e.g. via brainstorming, of any potential ethical impacts at the beginning of the software development process (e.g. when defining the software requirements).	*"If you use that [machine learning algorithms] then I think you need to address it [potential ethical impacts]. When that would be exactly…? Actually, after the requirements. Because with the requirements you only formulate what you want to have, not with what".* *(Interviewee 6)*	-
Decision about necessity to apply ethical development	Evaluation, whether the identified ethical impacts are critical enough to justify extra effort of ethical software development.	*"You would need some kind of check system to evaluate that. Is it a potential risk? [...] who could be a potential client? Who could benefit from it?"* *(Interviewee 7)*	*Spiekermann and Winkler (2020)*
Project division into ethically critical and non-critical components	Division of (especially large) software development projects into components that require ethical software development and those that do not. This allows development of noncritical components in the usual way.	*"[...] if you have a big project, maybe only one part is morally relevant [...], why should you let the rest be influenced by that part? It needs a completely different treatment than the rest". (Interviewee 4)*	-
Unchanged coding phase	Every software development project contains a phase of developing the actual software that should remain unchanged.	*"But if you then go into the development phase, then I'm in such a tunnel that I write my code. It has to work then and if I questioned anything, I wouldn't write any more code". (Interviewee 2)*	*Cary et al. (2003)*
Design & implementation of concrete solutions to identified critical ethical impacts	If applicable, concrete and implementable solutions to the identified ethical impacts should be included in the software's design.	*"[...] what do I want from my product and what do I not want to be done with my product' and then actually try to solve it technically, if that is possible, of course. [...] if you say that this is a product for 12- to 16-year-olds, then you put in a lockage after 2h that they don't game too much. That would be a concrete technical solution". (Interviewee 4)*	*Aydemir and Dalpiaz (2018)* *Cary et al. (2003)* *Manders-Huits (2011)* *Winkler and Spiekermann (2018)*

Table 3.4-4: Expert recommendations on ethical software development

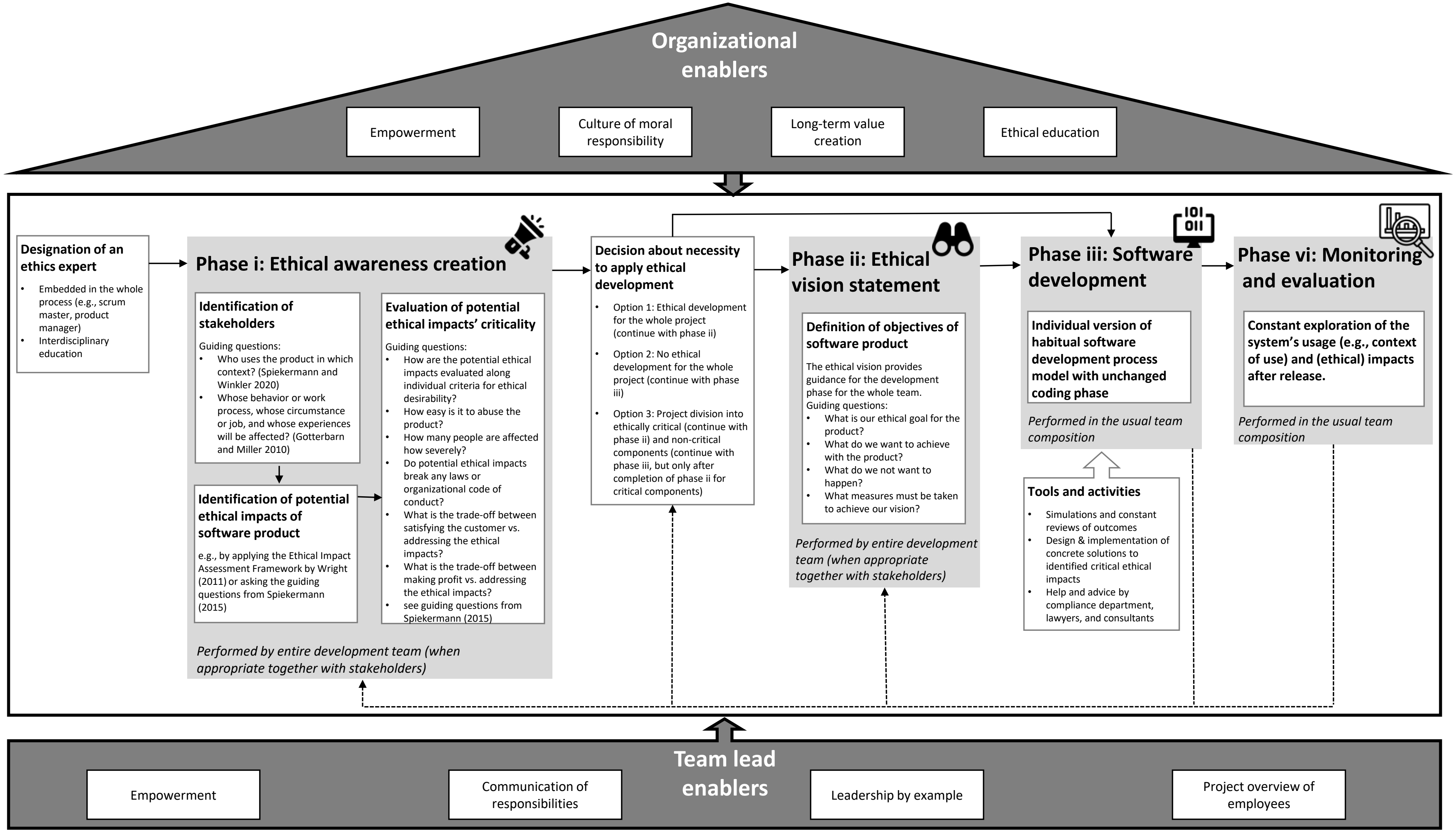
We define the following design principle:

(DP.6) An approach to ethical software development should incorporate experts' recommendations of what an ethical software development process should contain.

3.4.5 Design artifact description and design specifications

The design artifact is a process model which can be applied by software development teams independent of their choice of habitual software development process model. Our analysis reveals that ethical software development concerns not only the development process itself, but also the surrounding general conditions of the organization and the team. To account for this, the ESDP, depicted in Figure 3.4-1, contains two parts: First, a software development process with phase i to vi to guide software development teams through a development project. Second, the process is surrounded by organizational and team lead enablers, which are activities that, if implemented, motivate ethical software development.

Motivating factors (category 1) and expert recommendations (category 4) appear word by word in our model. The formulation of hindering factors (category 2) was first turned into positive and then incorporated into the artifact. Note that the hindering factor "lack of ethical awareness" founded the basis for whole phase i ("ethical awareness creation"). Procedural aspects (category 3) are adapted word by word, apart from the factor "Ethical considerations overarching all development process phases", which is reflected in our model by three explicit ethical phases (phase i, ii, and vi) and ethical enablers supporting the process. Furthermore, the factor "Taking the stakeholders' perspectives" is considered in phase i, especially in the first step ("identification of stakeholders") and the resulting identification of potential ethical impacts.



Organizational enablers

Empowerment

Culture of moral responsibility

Long-term value creation

Ethical education

Designation of an ethics expert

- Embedded in the whole process (e.g., scrum master, product manager)
- Interdisciplinary education

Phase i: Ethical awareness creation

Identification of stakeholders

Guiding questions:

- Who uses the product in which context? (Spiekermann and Winkler 2020)
- Whose behavior or work process, whose circumstance or job, and whose experiences will be affected? (Gotterbarn and Miller 2010)

Identification of potential ethical impacts of software product

e.g., by applying the Ethical Impact Assessment Framework by Wright (2011) or asking the guiding questions from Spiekermann (2015)

Evaluation of potential ethical impacts' criticality

Guiding questions:

- How are the potential ethical impacts evaluated along individual criteria for ethical desirability?
- How easy is it to abuse the product?
- How many people are affected how severely?
- Do potential ethical impacts break any laws or organizational code of conduct?
- What is the trade-off between satisfying the customer vs. addressing the ethical impacts?
- What is the trade-off between making profit vs. addressing the ethical impacts?
- see guiding questions from Spiekermann (2015)

Performed by entire development team (when appropriate together with stakeholders)

Decision about necessity to apply ethical development

- Option 1: Ethical development for the whole project (continue with phase ii)
- Option 2: No ethical development for the whole project (continue with phase iii)
- Option 3: Project division into ethically critical (continue with phase ii) and non-critical components (continue with phase iii, but only after completion of phase ii for critical components)

Phase ii: Ethical vision statement

Definition of objectives of software product

The ethical vision provides guidance for the development phase for the whole team.
Guiding questions:

- What is our ethical goal for the product?
- What do we want to achieve with the product?
- What do we not want to happen?
- What measures must be taken to achieve our vision?

Performed by entire development team (when appropriate together with stakeholders)

Phase iii: Software development

Individual version of habitual software development process model with unchanged coding phase

Performed in the usual team composition

Tools and activities

- Simulations and constant reviews of outcomes
- Design & implementation of concrete solutions to identified critical ethical impacts
- Help and advice by compliance department, lawyers, and consultants

Phase vi: Monitoring and evaluation

Constant exploration of the system's usage (e.g., context of use) and (ethical) impacts after release.

Performed in the usual team composition

Team lead enablers

Empowerment

Communication of responsibilities

Leadership by example

Project overview of employees

Figure 3.4-1: Ethical software development process model

The model starts with the designation of an expert for ethics. This person should be embedded in the whole development process and have an interdisciplinary education, including management, ethics, and technology. This is the first step to raise awareness among team-members for potential ethicality of the development process.

3.4.5.1 Phase i: Ethical awareness creation

Software products may possess no or neglectable ethical impacts and do not require ethical software development. However, the artifact should guide any software development team disregarding the nature of their product. Hence, phase i aims at raising awareness to the possibility of ethical impacts and empower software development teams to make an informed decision about whether to pursue ethical software development or not. Phase i consists of three process steps: Identification of stakeholders, identification of potential ethical impacts, and evaluation of their criticality. All process steps should be executed by the entire software development team to ensure that every member's ideas and concerns are heard, and everyone feels involved and committed to the software development team's decision. When appropriate, even stakeholders should be actively involved in phase i.

First, the team identifies major and minor stakeholders of the software product. Leading questions might be "Who uses the product in which context?" (Spiekermann and Winkler 2020) or "Whose behavior or work process, whose circumstance or job, and whose experiences will be affected?" (Gotterbarn and Miller 2010).

After, the team identifies the software product' potential ethical impacts. To approach this issue, we propose to follow Wright (2011) or, more value-oriented, Spiekermann (2015). The ethical impacts assessment framework for teams developing an information technology project introduced by Wright (2011) assists development teams to assess possible ethical impacts along four ethical principles (Beauchamp et al. 2001): Respect for autonomy, including issues of dignity and informed consent, non-maleficence, including issues of safety, social solidarity, isolation, and discrimination, beneficence, including issues of universal service, accessibility, value sensitive design, and sustainability, and justice, including issues of equality and fairness, as well as privacy and data protection. For each principle and its corresponding issues, the author proposes a set of questions to assess the product's ethical impact. For example, to address beneficence, Wright (2011) proposes to ask, "[w]ill the project provide a benefit to individuals? If so, how will individuals benefit from the project (or use of the technology or service)?" (p. 208). Spiekermann (2015) introduces three guiding questions that could equally serve to identify potential ethical impacts and even serve to evaluate their criticality in the next step of phase i: (1) What are all thinkable positive and negative consequences you can envision from the system's use for direct and indirect stakeholders? (2) What are the negative implications of the system for the character and/or personality of direct and indirect stakeholders (3) Which of the identified values and virtues would you consider as so important that you would want their protection to be recognized as a universal law? (Spiekermann 2015).

As a next step, the software development team assesses the identified ethical impacts regarding their criticality. The evaluation should be done along criteria that fit the organization's, project's, and software development team members' ethical values. Furthermore, the model proposes the following six questions (all extracted from our data analysis of the interviews) to guide the software development team through evaluating potential ethical impacts.

- How are the potential ethical impacts evaluated along individual criteria for ethical desirability?
- How easy is it to abuse the product?
- How many people are affected how severely?
- Do potential ethical impacts break any laws or organizational code of conduct?
- What is the trade-off between satisfying the customer vs. addressing the ethical impacts in further development?
- What is the trade-off between making profit vs. addressing ethical impacts in further development?

As mentioned before, the guiding questions from Spiekermann (2015) could equally be consulted in this step.

After phase i, the software development team decides whether the potential ethical impacts of the software product are sufficiently critical to require ethical software development through the remainder of the process. The findings of the previous evaluation of criticality enable the team to reach an informed decision. However, especially when considering large software development projects, it may be prudent to split it into ethically critical and non-critical components and address them differently. Depending on this decision, the ESDP next either enters phase ii (if further pursuit of ethical software development is required) or phase iii (if further pursuit of ethical software development is not required).

3.4.5.2 Phase ii: Ethical vision statement

If ethical software development is required, an ethical vision of the product goal should be defined by the entire software development team. The vision statement's aim is to guide the team through the development process and should clearly and concisely define why the software development team intends to develop the product and in how far ethics impacts the software's aim or requirements. To support formulation of an ethical vision statement, the model proposes the following four guiding questions, extracted from our data collection via interviews, to encourage intensive reflection.

- What is our ethical goal for the product?
- What do we want to achieve with the product?
- What do we not want to happen?
- What measures must be taken to achieve our vision?

The last question is especially important as it encourages the team to commit themselves to concrete actions to avoid losing sight of the vision.

3.4.5.3 Phase iii: Software development

After the team has defined an ethical vision statement, the ESDP enters phase iii. Alternatively, after phase i, the software development team decides to skip phase ii and directly enters phase iii. The ESDP's third phase allows software development teams to approach development using their habitual software development process models. This ensures that the development process fits the requirements of the project and context and the software development team members can work without having to adjust to a new and unfamiliar process. To support ethical considerations during phase iii, the model proposes the following three tools and activities, each extracted from interviews and literature (e.g., Wallnau (2018), Spiekermann and Winkler (2020), Aydemir und Dalpiaz (2018)):

- Simulations and constant reviews of outcomes during the development process, e.g., accessibility or privacy reviews
- Design and implementation of concrete solutions to identified critical ethical impacts
- Help and advice by compliance department, lawyers, and consultants

The ESDP design is iterative. Hence, even if the software development team initially decides to skip phase ii, the model encourages to challenge the conclusions of phases i and ii on a regular basis (e.g., every 3 months, at the beginning of a new Scrum sprint or after every phase of the spiral model). This is important to secure adaptability to unexpected changes during the software development process and to ensure the consideration of all potential ethical impacts, even if they are not apparent at the beginning of the software development process.

3.4.5.4 Phase iv: Monitoring and evaluation

It is important to constantly monitor and evaluate the software after release to ensure ethical usage for the whole lifecycle of a software product. Therefore, the team must observe the context of use, as a new context might bring new ethical challenges. Also, independently of the context, it is vital to constantly explore if unforeseen ethical challenges arise. If so, the software has to be adjusted to the new challenges. Minor changes can be directly considered in phase iii or via minor changes of the ethical vision statement in phase ii. Major challenges will lead the team back to phase i e.g., in order to identify new stakeholders and reassess ethical impacts' criticality.

3.4.5.5 Organizational and team lead enablers

Enablers build a leadership on organizational and project level that foster motivating factors and help overcoming hindering factors.

The artifact proposes four organizational enablers:

First, the organization needs to ensure that its employees feel empowered to participate in the organization's ethical strategy and goals. This will further increase the employees' individual feelings of responsibility and commit them to the organizational culture. Second, build and emphasize an organizational culture of moral responsibility, for example by introducing an organizational code of conduct. Third, organizations should clearly focus on long-term value creation and communicate the (long-term) economic benefits of ethical conduct and software development. Fourth, educating all employees and especially the software engineering professionals on their responsibility concerning software products' ethical impacts. For example, this can be achieved by means of trainings or workshops and should be an integral feature of an organization's employee development scheme.

On the software development project level, the artifact proposes four team lead enablers to foster ethical software development:

First, team leads need to empower their team members on the project level analogously to how the organization should empower its employees on an organizational level. Second, team leads need to clearly communicate the responsibility each team member bears for the software product under development. Third, team leads need to lead by example and emphasize the importance of ethical software development to the team members. Fourth, especially in large software development projects, team leads should ensure that every team member has enough overview of the project scope to be able to influence the project's outcome.

3.4.6 Validation of the design specification

As described in section 2, we first apply feature comparison against competing artifacts and then turn to further expert interviews for Eval 2.

3.4.6.1 Feature comparison and competing artifacts

We first apply the method of feature comparison (Venable et al. 2012). To validate whether the model's design specifications suitably addresses the research aim, we discuss its characteristics against our design objectives and design principles. As competing artifacts, we select the prescriptive design approaches to ethical software development by Cary et al. (2003), Lurie and Mark (2015), Aydemir and Dalpiaz (2018), and Spiekermann and Winkler (2020), introduced in section 1 which, to the best of our knowledge, comprise all existing approaches to ethical software development. The feature comparison confirms that our model fulfils all design objectives from a stand-alone perspective. For a detailed discussion, see Table 19

	ESDP (this paper)	Cary et al. (2003)	Lurie and Mark (2015)	Aydemir and Dalpiaz (2018)	Spiekermann and Winkler (2020)[17]
Summary of the approach to ethical software development	*The ESDP can be applied to any software product, includes factors hindering and motivating ethical software development and allows the team to apply their choice of software development process model. It includes current practices of ethical software development and implements industry experts' recommendations concerning the ideal ethical software development process.*	*The authors propose to embed ethical requirements into the SDLC. Their approach supports software development teams that apply the SDLC to develop data mining software. Motivating and hindering factors are partially considered. It is unclear to what extent existing practices and expert recommendations are considered.*	*The ethical-driven software development framework proposes a set of yes/no questions to overcome a lack of awareness to ethical implication during the SDLC of any software product. It does not consider motivating factors or current practices of ethical software development. The authors do not provide insights into their research method.*	*The ethics-aware software engineering framework is designed to raise awareness to potential ethical impacts during the development process. It can be applied to any software product. Motivating factors are considered through four enablers. It remains unclear if current ethical software development practices or expert recommendations are incorporated.*	*The methodological overview for ethics by design suggests 16 recommendations and 12 requirements for VBE. VBE is split into an ethical exploration phase (e.g. value elicitation, prioritization, and requirement identification) and an ethically aligned design phase (e.g. risk assessment, system development).*
(DO.1) The approach should be applicable to any kind of software development process model.	Phase iii allows the software development team to apply their choice of software development process.	The approach is customized to the original version of the SDLC.	The approach is customized to the original version of the SDLC.	The approach is agnostic of software development process models.	No restriction concerning the development model is named.
(DO.2) The approach should guide software development teams to consider ethical impacts of software products during development.	The model can be applied to the development of any software product, even for products that do not have obvious ethical impacts.	The approach is aimed specifically at data mining software products.	The framework can be applied to the development of any software product.	The framework can be applied to the development of any software product.	VBE can be applied when creating new technologies or to existing technologies. There exists a light risk-based process for less critical cases.
(DP.3) The approach should incorporate motivating factors of ethical software development.	The ESDP applies factors motivating ethical software development through phase ii (Ethical vision statement) and through the organizational and team lead enablers which transform the motivating factors into practices.	The approach includes requirements that aim at defining a product goal as well as incorporating organizational factors, such as the culture of moral responsibility (e.g., phases requirements analysis and design).	The proposed yes/no questions are aimed at raising awareness to ethical implications rather than factors that motivate ethical software development.	The approach includes four enablers that motivate and facilitate ethics-aware software engineering.	The approach names current challenges that are partly formulated as factors that, if applied, motivate ethical software development.
(DP.4) The approach should be designed to overcome the factors hindering ethical software development.	By introducing the organizational and team lead enablers, as well as the application of phase i, the ESDP is specifically designed to overcome the identified hindering factors.	The approach is designed to overcome a lack of ethical awareness.	The ethical-driven software development framework is designed to overcome a lack of awareness of ethical implications.	The ethics-aware software engineering framework is designed to overcome a lack of awareness concerning ethical impacts of software products.	The approach names current challenges and partly provides recommendations on how to overcome those.
(DP.5) The approach should incorporate procedural good practices of ethical software development processes.	Phase ii aims at the definition of a product goal; through the iterative process design, ethical considerations are applied throughout the entire software development process; simulations and constant reviews are proposed to apply during phase iii.	The ethical requirements embedded in the SDLC are based on pre-defined ethical principles. However, it is unclear how the authors define these principles and if current software development practices are considered.	The ethical-driven software development framework is dedicated to integrating awareness and understanding of ethical implications as an integral component of the SDLC and does not include other measures of ethical software development.	It is not apparent how the authors developed their framework and whether it is derived from current software development practices.	It is not apparent what aspects are drawn from the gained insights of IEEE P7000, which likely includes good practices.
(DP.6) The approach should incorporate experts' recommendations of what an ethical software development process should contain.	Phases i, ii, and iii, as well as the proposed tools include expert recommendations for ethical software development.	The ethical requirements embedded in the SDLC are based on pre-defined ethical principles. However, it is unclear how the authors define these principles and whether expert recommendation are considered.	It is unclear how the authors developed the ethical-driven software development's questionnaire and whether it is derived from current practices or from expert recommendations.	It is not apparent how the authors developed their framework and whether it is derived from expert recommendations for ethical software development.	It is not apparent what aspects are drawn from the gained insights of IEEE P7000, which likely includes experts' recommendations.

Table 3.4-4: Feature comparison with competing artifacts

Legend: Dark-grey: design objective or principle is fulfilled; light-grey: Design objective or principle is partly fulfilled; white: design objective or principle is not fulfilled or fulfillment cannot be judged

[17] Relates to IEEE P7000

3.4.6.2 Evaluation interviews with industry experts

To validate the model's design specifications from a naturalistic perspective, we conduct two additional evaluation interviews with industry experts (interviewee 4 from the first round of interviews and interviewee 8) which intend to assess the model's understandability, completeness, and real-world fidelity. In the first interview, we present the expert our ESDP "prototype" as a result of the first round of interviews and – walking him through the model in detail – ask for feedback on the artifact's design specifications regarding its real-world fidelity. In contrast to this, the second evaluation interview aims at testing the artifact's real-world fidelity by simulating the ESDP's implementation in an exemplary software development process. The expert simulated the ESDP by hypothetically recapitulating one of his real past software development projects and how the application of the ESDP would have affected the development process.

Interviewee 4 stresses the importance of a strong culture of moral responsibility as one of the major organizational enablers of ethical software development. The feeling of moral responsibility needs to be deeply rooted in the organization's identity to ensure that everyone "works as one". Further, he explicitly understands and agrees with the structure of the artifact. He especially approves of phase i, as he states that ethical awareness creation needs to happen at the very beginning of the development process to reduce the danger of unexpected repercussions. He also agrees with the approach of phase iii, and affirms that in his experience, every software development process is different and adapted to the software development team's individual needs. He notes the importance of educating employees as an organizational enabler and proposes to engage external coaches and consultants. The expert further remarks that software development team leads should also be responsible for educating their members on the special needs of selected projects. Lastly, the expert emphasized the importance of a feedback loop between the phases of the ESDP as he is currently staffed on a project which has lasted for three years. Consequently, the original scope of the project as well as the product requirements have considerably changed over the course of this time. There needs to be the possibility and an incentive to always dial back to phases i and ii throughout the software development process to evaluate whether the decisions made at the beginning of the project are still up to date. The original version of the ESDP presented to interviewee 4 did not contain such a feedback loop between the ESDP phases. After careful consideration, we incorporated the feedback loop from phase iii to phases i and ii.

The second evaluation interview simulates the ESDP along one of interviewee 8’s real past software development projects. The project’s scope was the development of a medical software application which interviewee 8 and his team developed for a customer. The interviewee was the lead software developer in the team at a young Munich-based R&D organization specialized in projects on emerging technologies. The selected use case is fitting to the ESDP as halfway through the project the customer unexpectedly added a request to track sensitive customer data via the application. This forced the software development team to weigh the trade-off between satisfying the customer’s wishes against following their own moral values. The expert states that if they had followed the ESDP, the customer might have had to address their request for data tracking already at the beginning of the project. As a subsequent step, the software development team would then have evaluated the criticality of the ethical impacts stemming from the data tracking. He proposes to rate the impacts along two criteria: First, how many people will be affected by the data tracking and second, how easily the collected data can be used to build detailed user profiles. He reasons that to provide the right evaluation criteria, one should always put her- or himself in the users’ shoes and ask themselves what the users would find bothering. In our model, we even go beyond the user’s perspective as we incorporate major and minor stakeholders. He approves of phase iii and would have developed the software application using the habitual software development process model in the use case. However, the expert remarks on phase ii as he considers other past software development projects: He argues that a software development team which is committed to an ethical vision statement may find itself in a scenario where it would have to find a compromise with the customer or else risk losing the project.

The experts both approve of the ESDP as a sensible approach to ethical software development. They agree that there currently exists a need in practice to guide software development teams to consider the ethical impacts of their software products during development. They both approve understandability of the model and apart from the feedback loop from phase iii to phases i and ii, which we added to our ESDP, they did no bring up further missing aspects, assuming completeness for our artifact. Both interviewees, however, remark at the end of their interviews that in their opinion the implementation and establishment of a new software development process model in existing organizational structures might pose a challenge, especially when considering large organizations, but nevertheless assume that the ESDP is applicable in practice.

3.4.7 Discussion

The approach to ethical software development outlined here shows that there are, among others, three main aspects to consider during ethical software development: First, the traditional software development process model should be encircled by additional steps to ensure sufficient ethical considerations before and after the actual development phase, as for instance designation of an ethics expert at the very beginning of the process. Second, the team lead has special responsibility to care for ethical development that should be supported by team lead enablers, as for instance leadership by example. Third, enablers on organizational level also facilitate ethical development, for example, via a culture of moral responsibility.

The ESDP shows that ethical software development is not a single optional step before or during the development phase that can be integrated if needed. Ethical software development requires a holistic view of the project and its goals from the very beginning. It is desirable that all software projects, and especially AI-software projects, where the future decisions of the AI are not always understandable for users (and neither predictable for developers), the development team should always follow the ESDP to ensure consideration of potential ethical impacts (Allen 2010; Russell et al. 2016). Therefore, the organization as well as the team-lead has the responsibility to set up enablers for ethical software development. Nevertheless, it is key that the team works openly together on each step of the ethical software development process model, from the designation of an ethical expert until the review and evaluation after release. For staffing, this means that software developers not only require good programming skills but should have interdisciplinary perspectives to bring into the process. Another far-reaching alteration from the traditional approach is the active inclusion of minor and major stakeholders. This includes not only the future users of the program, that are nowadays now and then involved into the process but requires a broader understanding of all stakeholders potentially affected by the product. In phase i, a useful evaluation of the ethical criticality requires an intense engagement with each stakeholder-group. For a proper basis for decision-making about pursuing ethical software development or not (after phase i), it is important to structurally consider all stakeholders, even if this step might be lengthy and rich in debate.

From the practical side, the ESDP certainly requires more interdisciplinary perspectives, more time, and more discussion than traditional software development. Nevertheless, it is of utmost importance that product development includes the additional steps and thoughts to ensure that our (future) software products do not make decisions against our ethical values (Allen et al.

2006). To the best of our knowledge, our artifact is the first to provide concrete guidance with leading questions for software development teams.

From the theoretical side, we showed that there exist few frameworks in literature that treat some facets of ethical software development (Aydemir and Dalpiaz 2018; Cary et al. 2003; Lurie and Mark 2015; Spiekermann and Winkler 2020), but no existing approach fulfills all design objectives and principles. Consequently, we contribute to the literature about ethical software development in three ways: First, we provide an overview of existing ideas and approaches about ethical software development. Second, we enrich the existing literature via the data analysis of interviews with industry experts. Third, we brought those findings together into our integrated process model of ethical software development that guides software development teams.

With these three aspects, this paper contributes to a theory of design and action, as we say "how to do something" with "prescriptions […] for constructing" software (Gregor 2006, p. 620). Within the three-level categorization introduced by Gregor and Hevner (2013), we therefore contribute to a level 2 nascent design theory. Specifically, the key design artifact is the process model depicted in Figure 3.4-1 and detailed in the text. This key artifact builds on other artifacts, namely the constructs that are introduced in sections 2 and 4 and summarized in the model, and the design principles 1-4 presented in section 4.

In the following, we briefly summarize our contribution to theory based on Gregor and Jones (2007), who identified the following core components of a information systems design theory: purpose and scope, constructs, principle of form and function, artifact mutability, testable propositions and justificatory knowledge. Our process model is applicable to all software development projects and is especially important for projects including AI-components as here, ethical impacts are particularly difficult to foresee. The purpose of our process model is to improve software development in terms of considering the ethical impacts of software. Our constructs are the factors described in Table 14 to Table 18, summarized in our process model in Figure 3.4-1. In line, the principle of form and function, meaning the "blueprint or architecture that describes an IS artifact" (Gregor and Jones 2007, p. 322), is represented by our ESDP, including all constructs. Regarding artifact mutability, our process model can be applied to all kind of software development processes, as for instance SCRUM or waterfall model. For this, the process model as presented here needs to be adapted to the specific software development process. Our propositions are that our ESDP fulfills design objectives 1 and 2 and that it does so better than extant models. Furthermore, our theory helps to reduce ethical issues in software. For gaining justificatory knowledge, we conducted a broad literature research and conducted interviews with industry experts.

As any research endeavor, our research is beset with limitations. First, the number of interviews is relatively small and limited to Germany. Further research might significantly increase the number of interviews for a richer data base, which might open up opportunities for cultural comparison, comparison between different sizes of companies, or between different project scopes.

Furthermore, our literature research focused on scientific literature, leaving aside grey literature. Especially for such a dynamic field as ethics in AI, sources next to scientific journals might reveal interesting results that are on the pulse of time.

Additionally, our research did not explicitly focus on the users' perspective. The ESDP includes stakeholders of the product, which should include users among others, but no special weight is lied on them. Hence, user centered design models could be compared to the ESDP and further research could extend or adapt the approach.

For future research, the model could be tested with real-life software development projects to identify potential obstacles in implementation that might expand our model. Furthermore, we expect real-life implementation of the model to bring up interesting, new thoughts. One aspect that could be analyzed in future research was brought up from interviewee 8 arguing that even in critical cases, refusing to accept a project offer would not be an option for him because other organizations would surely do it anyway. Only if he and his team take on the project and try to persuade the customer to review the product's requirements, they can influence the outcome and try to align the product to their ethical values. Following this thought, potential options and its consequences of each part of the model could be interesting to study in future research.

Furthermore, future research could examine the impact on time and costs of a software development project that follows the ESDP, compared to other, more traditional approaches. This might help organizations to make informed decisions about potential inclusion of ethical considerations.

For future research, we expect interesting results from comparison of different kind of software development projects, for instance considering the type of technology implemented (e.g., using black box machine-learning models) and based on size of the project, size of the development team, or size of the organization. Especially team lead and organizational enablers might vary depending on these variables. Research could analyze if there is an advantageous team, project, or organizational size for ethical development.

3.4.8 Conclusion

In order to prevent negative ethical consequences, software development teams need structural guidance to consider ethical impacts during the software development process. The aim of this research was to develop an ethical software development process model that provides guidance on how software development teams should incorporate the software product's potential ethical impacts during the design and development process. To address this research objective, we develop an nascent design theory (Gregor and Hevner 2013). Based on literature and qualitative, semi-structured interviews with industry experts, we derived design objectives and design principles and developed our design artifact, the ESDP. Next to team-lead and organizational enablers, our process model introduces four phases (ethical awareness creation, ethical vision statement, software development, monitoring and evaluation) and, based on literature and interviews, provides guiding questions and tools that support development teams when executing the respective phase. For evaluation, we compared the features of our model to competing artifacts and conducted evaluation interviews with industry experts. Our holistic process model of ethical software development should be used for all kind of software to actively engage with (potential) ethical impacts, obvious or not. Furthermore, single steps of our artifact can be the starting point for future research to dive deeper into the field of research which is becoming increasingly important.

References

Al-A'ali, M. (2008). Computer ethics for the computer professional from an Islamic point of view. *Journal of Information, Communication and Ethics in Society*, 6(1), 28–45.

Allen, C. (2010). Artificial life, artificial agents, virtual realities: technologies of autonomous agency. In L. Floridi (Ed.), *The Cambridge handbook of information and computer ethics* (pp. 219–233). Cambridge, UK: Cambridge University Press.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3), 149–155.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3), 251–261.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, 21(4), 12–17.

Anderson, M., & Anderson, S. (2007). Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 28(4), 15–26.

Anderson, M., Anderson, S., & Armen, C. (2006). An approach to computing ethics. *IEEE Intelligent Systems*, 21(4), 56–63.

André, E., Bayer, S., Benke, I., Benlian, A., Cummins, N., Gimpel, H., . . . Weber, K. (2019). Humane Anthropomorphic Agents: The Quest for the Outcome Measure. *AIS SIGPRAG Pre-ICIS Workshop*.

Arquilla, J. (2010). *Conflict, security and computer ethics*. In L. Floridi (Ed.), The Cambridge handbook of information and computer ethics (pp. 133–148). Cambridge, UK: Cambridge University Press.

Association for Computing Machinery. (2018). *ACM Code of Ethics and Professional Conduct*. Retrieved from https://www.acm.org/code-of-ethics

Aydemir, F. B., & Dalpiaz, F. (2018). A roadmap for ethics-aware software engineering. In *Proceedings of the International Conference on Software Engineering*, ICSE 2018 (pp. 15–21).

Beauchamp, T., Childress, J., & others. (2001). *Principles of biomedical ethics*: Oxford University Press, USA.

Bello, P., & Bringsjord, S. (2013). On How to Build a Moral Machine. *Topoi*, 32(2), 251–266.

Bogosian, K. (2017). Implementation of Moral Uncertainty in Intelligent Machines. *Minds and machines*, 27(4), 591–608. doi:10.1007/s11023-017-9448-z

Brandenburg, S., & Minge, M. (2019). Epos – an instrument for the assessment of the ethical position in software development. *Theoretical Issues in Ergonomics Science*, 20(2), 153–165. doi:10.1080/1463922X.2018.1491072

Brey, P. (2010). *Values in technology and disclosive computer ethics*. In L. Floridi (Ed.), The Cambridge handbook of information and computer ethics. Cambridge, UK: Cambridge University Press. Retrieved from 41-58

British Computer Society. (2020). *BCS, the chartered institute for IT - code of conduct for BCS members*. Retrieved from https://cdn.bcs.org/bcs-org-media/2211/bcs-code-of-conduct.pdf

Cary, C., Wen, H. J., & Mahatanankoon, P. (2003). Data mining: Consumer privacy, ethical policy, and systems development practices. *Human Systems Management*, 22(4), 157–168.

Cervantes, J.-A., Rodríguez, L.-F., López, S., Ramos, F., & Robles, F. (2016). Autonomous Agents and Ethical Decision-Making. *Cognitive Computation*, 8(2), 278–296. doi:10.1007/s12559-015-9362-8

Collins, R., & Miller, K. (1994). How good is good enough? An ethical analysis of software construction and use. *Communications of the ACM*, 37(1), 81–92. doi:10.1145/175222.175229

Copp, D. (Ed.). (2006). *The Oxford handbook of ethical theory:* Oxford, UK: Oxford University Press.

Corbin, J., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1), 3–21.

Davis, A., Bersoff, E., & Comer, E. (1988). A strategy for comparing alternative software development life cycle models. *IEEE Transactions on Software Engineering*, 14(10), 1453–1461. doi:10.1109/32.6190

Davis, J., & Nathan, L. (Eds.). (2015). *Value Sensitive Design: Applications, Adaptations, and Critiques*: Springer.

Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., . . . Wildt, T. de. *Ethics by Design: Necessity or Curse?* In Furman, Marchant et al. 2018 – AIES'18 (pp. 60–66).

Dodig C., & Çürüklü, B. (2012). Robots: Ethical by design. *Ethics and Information Technology*, 14(1), 61–71. doi:10.1007/s10676-011-9278-2

European Comission. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Floridi, L. (Ed.). (2010). *The Cambridge handbook of information and computer ethics:* Cambridge, UK: Cambridge University Press.

Floridi, L., & Sanders, J. (2004). On the morality of artificial agents. *Minds and machines*, 14(3), 349–379.

Fossa, F. (2018). Artificial moral agents: moral mentors or sensible tools? *Ethics and Information Technology*, 20(2), 115–126. doi:10.1007/s10676-018-9451-y

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. *Early engagement and new technologies: Opening up the laboratory*, 55–95.

Fritz, A., Brandt, W., Gimpel, H., & Bayer, S. (2020). Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI). *De Ethica: A Journal of Philosophical, Theological and Applied Ethics*, 6(1), pp.3-22.

Génova, G., González, M., & Fraga, A. (2007). Ethical education in software engineering: Responsibility in the production of complex systems. *Science and Engineering Ethics*, 13(4), 505–522.

Gips, J. (1995). Towards the Ethical Robot. Android Epistemology: Cambridge, MIT Press.

Gotterbarn, D., & Miller, K. (2010). Unmasking your software’s ethical risks. *IEEE Software*, 27(1), 12–13. doi:10.1109/MS.2010.23

Gotterbarn, D., Miller, K., & Rogerson, S. (1997). Software engineering code of ethics. *Communications of the ACM*, 40(11), 110–118.

Gregor, S. (2006). The Nature of Theory in Information Systems. *Management Information Systems Quaterly*, 30(3), 611–642.

Gregor, S., & Hevner, A. (2013). Positioning and presenting design science research for maximum impact. *Management Information Systems Quaterly*, 337–355.

Gregor, S., & Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information systems*, 8(5), 312–335.

Hameed, S. A. (2009). Software engineering ethical principles based on Islamic values. *Journal of Software*, 4(6), 563–570.

Hameed, S. A., Al-Khateeb, K., & Mutaz, Z. (2010). Software engineer Islamic ethics: An interactive web-based model. In *International Conference on Computer and Communication Engineering*, ICCCE'10 . Kuala Lumpur.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 11(1), 19–29. doi:10.1007/s10676-008-9167-5

Himma, K. E., & Tavani, H. T. (Eds.). (2008). *The handbook of information and computer ethics:* Hoboken, N.J.: Wiley.

Honarvar, A., & Ghasem-Aghaee, N. (2009). An artificial neural network approach for creating an ethical artificial agent. In IEEE *International Symposium on Computational Intelligence in Robotics and Automation*, CIRA 2009 (pp. 290–295).

IEEE P7000 Working Group. (2020). IEEE P7000 Working Group. Retrieved from https://sagroups.ieee.org/7000/

Jia, J., & Xin, J. (2018). Integration of ethics issues into software engineering management education. In *ACM International Conference Proceeding Series 2018* (pp. 33–38). Association for Computing Machinery.

Johnson, D. (1985). *Computer Ethics*: New Jersey: Englewood Cliffs.

Johnson, D. (2004). *Computer ethics.* The Blackwell guide to the philosophy of computing and information, 65–75.

Judy, K. H. (2009). Agile principles and ethical conduct. In R. H. Sprague Jr. (Ed.), Proceedings of the *42nd Annual Hawaii International Conference on System Sciences,* HICSS 2009 . Waikoloa, HI.

Karim, N., Ammar, F. A., & Aziz, R. (2017). Ethical Software: Integrating Code of Ethics into Software Development Life Cycle. In *International Conference on Computer and*

Applications, ICCA 2017 (pp. 290–298). Institute of Electrical and Electronics Engineers Inc.

Krcmar, H. (2015). *Informationsmanagement* (6th ed.): Berlin Heidelberg: Springer.

Lehnert, M., Linhart, A., & Röglinger, M. (2016). Value-based process project portfolio management: integrated planning of BPM capability development and process improvement. *Business Research*, 9(2), 377–419.

Lonchamp, J. (1993). A structured conceptual and terminological framework for software process engineering. In *Proceedings of the Second International Conference on the Software Process-Continuous Software Process Improvement*, ICSE 1993 (pp. 41–53).

Lurie, Y., & Mark, S. (2015). Professional Ethics of Software Engineers: An Ethical Framework. *Science and Engineering Ethics*, 22(2), 417–434. doi:10.1007/s11948-015-9665-x

Manders-Huits, N. (2011). What values in design? The challenge of incorporating moral values into design. *Science and Engineering Ethics*, 17(2), 271–287. doi:10.1007/s11948-010-9198-2

McNamara, A., Smith, J., Murphy-Hill, E., & Garci A., Pasareanu C.S., Leavens G.T. (2018). Does ACM's code of ethics change ethical decision making in software development? In *Proceedings of the 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2018 (pp. 729–733).

Microsoft. (2020). *Microsoft AI principles*. Retrieved from https://www.microsoft.com/en-us/ai/our-approach-to-ai

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook*: Thousand Oaks, California: Sage Publications.

Mitcham, C. (1995). Computers, information and ethics: A review of issues and literature. *Science and Engineering Ethics*, 1(2), 113–132.

Moor, J. (1985). What is computer ethics? *Metaphilosophy*, 16(4), 266–275.

Moor, J. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21.

Myers, M., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and organization*, 17(1), 2–26.

OECD. (2020). *OECD Principles on AI.* Retrieved from https://www.oecd.org/going-digital/ai/principles/

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77.

Pries-Heje, J., Baskerville, R., & Venable, J. (2008). Strategies for Design Science Research Evaluation. In *Proceedings of the European Conference on Information Systems* (pp. 255–266).

Rashid, A., Moore, K., May-Chahal, C., & Chitchyan, R. (2015). Managing Emergent Ethical Concerns for Software Engineering in Society. *Proceedings - International Conference on Software Engineering*, 2, 523–526. doi:10.1109/ICSE.2015.187

Rubin, H., & Rubin, I. S. (2011). *Qualitative interviewing: The art of hearing data*: Thousand Oaks, California: Sage Publications.

Ruparelia, N. (2010). Software development lifecycle models. *ACM SIGSOFT Software Engineering Notes*, 35(3), 8–13.

Russell, S., Norvig, P., Davis, E., & Edwards, D. (2016). *Artificial intelligence: A modern approach* (Third edition, Global edition). Always learning. Boston: Pearson.

Singer, P. (1986). *Applied ethics*: Oxford: Oxford University Press.

Slaughter, S. A., Levine, L., Ramesh, B., Pries-Heje, J., & Baskerville, R. (2006). Aligning software processes with strategy. *Management Information Systems*, 30(4), 891–918.

Sonnenberg, C., & vom Brocke, J. (2012). *Evaluations in the science of the artificial - reconsidering the build-evaluate pattern in design science research*. In K. Peffers & B. Kuechler (Eds.), Design Science Research in Information Systems -Advances in Theory and Practice (pp. 71–83). Berlin Heidelberg: Springer.

Spiekermann, S. (2015). *Ethical IT innovation: A value-based system design approach*: CRC Press.

Spiekermann, S., & Winkler, T. (2020). *Value-based Engineering for Ethics by Design*. preprint arXiv:2004.13676. Retrieved from https://arxiv.org/abs/2004.13676

Spinellis, D. (2017). The Social Responsibility of Software Development. *IEEE Software*, 34(2), 4–6. doi:10.1109/MS.2017.48

Stahl, B. C. (2010). *Social issues in computer ethics*. In L. Floridi (Ed.), The Cambridge handbook of information and computer ethics (pp. 101–115). Cambridge, UK: Cambridge University Press.

Sullins, J. (2010). *Rights and computer ethics*. In L. Floridi (Ed.), The Cambridge handbook of information and computer ethics (pp. 116–132). Cambridge, UK: Cambridge University Press.

Thomson, A. J., & Schmoldt, D. (2001). Ethics in computer software design and development. *Computers and Electronics in Agriculture*, 30(1-3), 85–102.

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A comprehensive framework for evaluation in design science research. *International Conference on Design Science*, 423–438.

Wallnau, K. C. (2018). Safety from ethical hazards: Prospects for a contribution from software engineering. *Proceedings - International Conference on Software Engineering*, 15–17. doi:10.1145/3195555.3195569

Wiegel, V. (2006). Building blocks for artificial moral agents. *Proceedings of Artificial Life*, 1–4.

Wiegel, V. (2010). *The ethics of IT-artefacts*. In L. Floridi (Ed.), The Cambridge handbook of information and computer ethics (pp. 201–218). Cambridge, UK: Cambridge University Press.

Wiltshire, T. J. (2015). A Prospective Framework for the Design of Ideal Artificial Moral Agents: Insights from the Science of Heroism in Humans. *Minds and machines*, 25(1), 57–71. doi:10.1007/s11023-015-9361-2

Winkler, T., & Spiekermann, S. (2018). Twenty years of value sensitive design: a review of methodological practices in VSD projects. *Ethics and Information Technology*. doi:10.1007/s10676-018-9476-2

Wright, D. (2011). A framework for the ethical impact assessment of information technology. *Ethics and Information Technology*, 13(3), 199–226.

Appendix

Appendix 3.4-A: List of interviewees

ID	Current position	Description of employing organization	Years of work experience
Interviewee 1	Software engineer	American multi-national technology organization. Number of employees: >100,000. City of employment: Paris, France.	1.5
Interviewee 2	Software engineer	Self-employed. Newly founded technology startup focusing on VR. Number of employees: <10. City of employment: Munich, Germany.	4
Interviewee 3	Software engineer, co-founder, CTO	Startup in the health care sector. Number of employees <10. City of employment: Munich, Germany.	7
Interviewee 4	Software engineer	German multinational automotive organization. Number of employees: >100,000. City of employment: Munich, Germany.	8
Interviewee 5	Software product owner, no technical background	Digital branch of a German organization offering, among others, experience vouchers. Number of employees: ca 300. City of employment: Munich, Germany.	7
Interviewee 6	Software engineer	Self-employed.	3
Interviewee 7	Software product owner, no technical background	Startup developing hardware and software to map, navigate and digitize the indoors. Number of employees: 100-250. City of employment: Munich, Germany.	13
Interviewee 8	Lead software engineer	R&D organization specialized in projects on emerging technologies. Number of employees: <50. City of employment: Munich, Germany.	Unknown

4 General discussion and conclusion

The following sections present the results and implications in section 4.1, limitations and suggestions for future research in section 4.2, and concluding thoughts in section 4.3.

4.1 Summary of results and implications

This dissertation focuses on opportunities, and in particular, challenges of the emerging technologies IoT and AI along the socio-technical continuum from Sarker (2019). Throughout, it does so from a human-centered perspective. Section 4.1.1 summarizes the key findings of the research articles from chapter 2 of this dissertation so as to achieve a better understanding of IoT, whereas section 4.1.2 covers those for chapter 3 with its focus on AI.

4.1.1 Results and implications of chapter 2: Behind the scenes of the Internet of Things

Chapter 2 examines the opportunities and challenges of IoT. First, section 2.1 analyzes the opportunities that IoT offers customers in the context of commerce. This analysis results in 12 affordances of IoT devices in the customer buying process. In section 2.2, the second research paper sheds light on the challenges that come with IoT. Specifically, it focuses on IoT's ethical challenges, which are discussed in literature, to propose directions for further research.

Section 2.1 argues that IoT-commerce has come to complement e-commerce and m-commerce. It builds on Activity Theory as a theory for analyzing and explaining, combined with affordances, defined as "possibilities for goal-directed actions of goal-oriented actors with regards to an object" (Bayer et al., 2021). The research article follows a two-step approach of theory development, followed by validation. In the first step of the theory development, the affordances of e-commerce, m-commerce, and IoT-commerce are identified. A structured literature review in the AIS Senior Scholars' Basket of Eight, complemented by journals in the electronic commerce and marketing domain, and a structured literature search of the literature of computer science and electrical engineering, resulted in 180 articles, whereupon 49 were classified as relevant for further examination of affordances. Ultimately, 12 affordances were identified. These include aspects such as *electronic transactions, personalized services, proactive services, natural interactions*, or *automated customer processes*. For validation, a sample of 337 IoT devices was screened for devices that facilitate or influence the customers' buying process. In the process, the number of relevant devices was narrowed down to 35, and these were then divided into five groups of IoT devices, such as *voice assistants, smart resource management*, or *rental services*. With regard to each group, a structured evaluation of completeness and parsimony was conducted. Our results show that IoT has the potential to transform the buying process, which makes it a highly relevant research topic. Most of the discussed affordances originated in e- or m-commerce. Three of them, however, are unique to IoT-commerce. The research article contributes to theory by conceptualizing IoT-commerce and identifying, conceptualizing, and linking the 12 affordances of IoT-commerce to the customer buying process. In practical terms, it provides new knowledge for customers and companies. As far as customers are concerned, a deeper understanding of IoT's potential is expected to foster critical reflections about customers' (future) self-determination in the buying process. For companies, there are opportunities to take advantage of new customer-oriented business models.

As indicated, even studies of the opportunities afforded by IoT touch on certain adverse side effects of IoT, such as the limits it places on the self-determination of customers, but most studies do so in passing. This leads to section 2.2, which explicitly looks at the ethical challenges associated with IoT, abbreviated to IoT ethics, in reference to the common term ICT ethics (Bernd Carsten Stahl & Rogerson, 2009). After a structured literature review in leading IS journals and a search in journals that deal with information or business in combination with ethics, 36 articles were found to combine the keywords "IoT" and "ethics" or "moral". Out of these, 17 articles addressed at least one ethical challenge linked to IoT. From those articles, relevant phrases were extracted, which, after consolidation, resulted in 21 ethical issues of IoT, such as *objectification of humans*, *unclear responsibility & accountability*, *privacy threats*, *questionability of informed consent*, *endangered physical safety*, or *technostress*. The issues were differentiated into four categories: *metaphysics, digital world, data and machine learning, and physical device*. Each category comprises at least two ethical aspects, *data and machine learning* being the largest with ten ethical issues. Each of these issues is described individually and discussed with a comment on the state of current research as well as a note on the features of IoT that have the greatest bearing on the respective issue. Furthermore, each is illustrated with an application example. As the results show, certain issues are not exclusively related to IoT but also to other emerging technologies. Examples include *job insecurity* and *complexity & opaqueness*. In contrast, other issues play a specific role due to IoT's unique characteristics, such as *tracking and monitoring* or *technostress*. Meanwhile, *ubiquity* is the most common trigger of IoT's ethical issues, followed by *sensing and actuating capabilities*. Most research articles on this topic do not provide a detailed discussion of the respective issue, but instead merely name or enumerate them. The contribution of this paper lies in the fact that it structures and categorizes the literature on IoT ethics, reveals a significant lack of in-depth research, and illustrates the importance of the latter with regard to IoT-specific challenges.

What these results of the first two research articles indicate is that IoT has the potential to radically transform individuals' daily life, illustrated by the example of commerce. However, there are potential ethical issues associated with the ever-increasing use of IoT, which issues can arise in various areas of life, and which have not yet been sufficiently examined. A major challenge in research about IoT might be the huge diversity of IoT devices, features, and application contexts. The research article on IoT ethics approached the field of IoT holistically, by aiming to provide an overview if IoTs ethical issues. The other research article focused on a particular application context, namely commerce, and here, too, the results show that the considerable diversity of devices and features that potentially influence the buying process of retail customers.

Furthermore, our research might point at the difficulty of a clear distinction between IoT and other (emerging) technologies. To some extent, this was to be expected, since technology convergence makes such a distinction near enough impossible (Jeong et al., 2015). Nonetheless, our results in both research articles show that there are aspects more relevant to IoT than others. For instance, the ethical issue of *social & economic exclusion* is not related to IoT in particular, but rather to digitalization and the increasing spread of technologies in general, whereas the issue of *tracking and monitoring* is closely related to IoT and facilitated by it. Likewise, the affordance of *electronic transaction* is not specifically boosted by IoT commerce, as opposed to *natural interactions*. On the one hand, this might indicate the difficulty of keeping research exclusively focused on IoT. On the other hand, however, our research has shown that a nuanced distinction is not only possible but also conducive to a better understanding of IoT and its related opportunities and challenges. Once again, then, our work underlines the need for IoT-focused research.

4.1.2 Results and implications of chapter 3: Behind the scenes of Artificial Intelligence

Chapter 3 examines the emergent technology that is AI from a human-centered perspective. Section 3.1 analyzes the concerns of individuals, followed by an ethical argumentation about moral agency and responsibility in section 3.2. The concept of trust in AI systems is examined in section 3.3, followed by a framework for ethical software development that is presented in section 3.4.

Section 3.1 considers the algorithms underlying AI. A structured literature search and a qualitative content analysis of semi-structured interviews were conducted to reveal potential concerns about ADM. To include relevant articles from IS and other disciplines that may contribute to this research, such as engineering, law, and marketing, we searched in the databases ACM Digital Library, AIS Electronic Library, Science direct, EBSCOhost, JSTOR Library, SpringerLink, and ProQuest. Of the 175 results, 18 were classified as relevant. To identify any further concerns, 13 semi-structured interviews were conducted, and for the purpose of qualitative content analysis, all interviews were transcribed and then analyzed in several steps, such as building a coding frame, segmentation, trial coding, and modifying the coding frame, based on the software MAXQDA. In the process, 24 concerns about ADM were identified. These were divided into separate categories such as concerns inherent to *technology* (e.g., *breakdown of technology) data* (e.g., *insufficient or wrong data basis)* or *decision* (e.g., *omission of human decision factors*), *physical* concerns, *social* concerns, *career-related* concerns, or *resource-related* concerns. With regard to each of these concerns, a description, the literature sources, and/or the interview IDs are provided. Only two concerns identified in the literature could not be confirmed in the semi-structured interviews (*job loss* and *environmental harm*). Aside from those found in the literature, 11 additional concerns were added through analysis of the interviews. The resulting framework of concerns about the use of ADM shows that concerns that fall into the category *decision* are unique to ADM, whereas most others also apply to emerging technologies such as IoT or Blockchain. The research also revealed certain mitigating circumstances, for instance when no difference is perceived between ADM and human decision-making or when there is high transparency. Further positive findings of the interviews include the potential for time-saving and reduced subjectivity of ADM. For theoretical integration, the framework of concerns about the use of ADM was compared to the related theory of Karwatzki et al. (2017). As a contribution to the research field of the dark side of IS, the framework increases the understanding and structures the concerns that prevent usage of ADM technologies. It guides the anticipation and evaluation of obstacles that could get in the way of ADM applications fulfilling their potential for market success. Furthermore, organizations can use this framework to prevent concerns prior to the dissemination of ADM technology as the framework makes it easier to address relevant concerns from the very beginning and thus increase user trust. At the same time, users can equally benefit from the framework by using it to systematically form their own opinions about concerns that may or may not apply to them for particular ADM technologies.

Section 3.2 takes a deep dive into one common concern about AI – responsibility. The lack of transparency associated with AI systems, often referred to as black-box character, raises the question whether humans can be responsible and accountable for the actions of an AI system they do not understand. Potentially, responsibilities could be allocated along the entire value chain of AI system development, from the algorithm developers to the data scientists all the way to the user. The first insight our research article provides is that, based on the actor-network theory, non-humans can also be accorded the term agency or moral agency, meaning that action and agency is a process distributed between entities such as humans and technology. Based on this thought, the techno-centric model of Floridi (Floridi, 2016; Floridi & Sanders, 2001, 2004), the anthropocentric model of Johnson and Verdicchio (Johnson & Verdicchio, 2018), and the constructivist model of Verbeek (Verbeek, 2006, 2011, 2014, 2017) are analyzed in an attempt to answer the question of responsibility in human-computer interaction. For Floridi's model, we argue that it is not suitable and helpful for our research aim, as it does not contribute to a better understanding of the human-computer interaction but rather aims to argue why computers can be perceived as agents – but without further meaningful argumentation. Meanwhile, Johnson and Verdicchio have a very inclusive use of the term agency for humans, computer systems, and human-computer interaction. This makes a clear distinction between the responsibilities of those entities impossible, which limits its contribution to our research aim. Verbeek's and Verdicchio's understanding of human-computer interaction is comprehensive, but their descriptions of moral agency are circular, which again restricts the benefit to our research. In sum, the research showed that there are significant risks when attributing agency to computational behavior. Hence, a clear distinction between humans and computers is essential in order to achieve a cogent analysis of the ethical-normative structure of human-computer interaction.

To address the various concerns that individuals have about AI, such as its lack of transparency or its potential for discrimination, XAI provides a sensible starting point. Section 3.3 discusses the concept of trust in explainable AI decision support systems. After reviewing the general literature on trust and AI as well as the specific literature on trust in technology and AI, we formulate five hypotheses that delineate the path from *trusting beliefs* to *behavior*, including *trusting intention*, *expertise*, and *explanation*. The hypotheses are tested empirically in the context of chess games, drawing on data collected in an online experiment with a self-developed AI system. The domain-specific expertise of each participant is assessed with the Amsterdam Chess Test. In the experiment, participants play out three chess scenarios. After each move, the supporting AI suggests a different move, whereupon the participant is informed that this suggested move is superior to the one made by the participant. This suggestion is either not further explained by the supporting AI or explained by enumerating the scenarios that are most likely to ensue from the move made by the participant as opposed to those most likely to follow from the move suggested by the AI, for example, “If you keep the move, the following scenarios are most likely: Black Bishop from b4 to c3, which can be answered by White Pawn from b2 to c3. Instead of your move, I would suggest […]”. After this suggestion, participants communicate their behavior by clicking either that they want to keep their own move or instead want to make the move suggested by the supporting AI. We collected data from 100 participants and analyzed our model with partial least square structural equation modeling. Our results show that trusting beliefs have a significant positive impact on trusting intention, and trusting intention positively affects behavior. User expertise showed a significant negative impact on trusting beliefs and expertise moderated the effect of explanation on trusting intention. The hypothesis that an explanation has a direct effect on trusting intention and behavior could not be supported. Furthermore, the average rate at which the suggestion of the supporting AI was followed only amounted to 40%, even though participants attributed high competence to the AI. To sum up, then, section 3.3 of this dissertation builds on prior literature and adapts it to the context of AI, contributing to research on trust in AI. Furthermore, we analyze the relationship between trusting intention and behavior by measuring actual behavior. This is most often left aside in existing literature, but, our results show that it is important to measure behavior, rather than merely approximate it by intention. Practitioners may benefit from our research in as much as it helps them understand how users build trust in AI systems and shed light on the importance of considering the domain expertise of the AI system’s target group if the system is to realize its full success potential.

As this research article showed, it is crucial to deal with the concept of trust prior to the development of an AI system. This point is further explored in section 3.4, where we argue that ethical considerations that may notably influence trust must be addressed from the beginning of a software design and development process. We followed the six steps of design science research as proposed by Peffers et al. (2007) with problem identification, definition of the objectives for a solution, design and development, demonstration, evaluation, communication. With regard to justificatory knowledge, we conducted a thorough literature review to establish the theoretical background on software development, computer and machine ethics, and ethics in software development. Furthermore, seven semi-structured interviews with industry experts allowed us to derive further design principles. Each interview was transcribed and analyzed with the help of the software MAXQDA. The analysis resulted in the ethical software development process, consisting of the core of the *ethical software development process, organizational enablers* such as *empowerment* and *ethical education*, and *team lead enablers* such as *leadership by example* and *communication of responsibilities*. The core depicts the path from the phase of ethical *awareness creation*, including for instance *identification of stakeholders, ethical vision statement, software development, monitoring and evaluation*. With regard to each phase, we provide a description and guiding questions, if applicable. For evaluation, feature comparison against competing artifacts is applied, followed by evaluation interviews. Our model summarizes and structures prior literature on ethical software development. We supplemented this with insights from our interviews. Meanwhile, our comprehensive framework offers guidance on how to include ethical considerations throughout the entire software design and development process. Hence, we contribute to a theory of design and action, more precisely to a level 2 nascent design theory (Gregor & Hevner, 2013). From the practical side, we demonstrated that ethical software development has to be more interdisciplinary and include additional steps in the design and development process to prevent software from making decisions against our moral values (Allen et al., 2006).

In summary, it is worth noting that the four research articles that make up this chapter indicate that AI poses multi-faceted challenges, some of which will undoubtedly have to be investigated in more detail. Potential solutions are on the way but need a holistic view and profound examination of AI’s challenges in order to address them successfully.

Our research might point at the urgent need for more interdisciplinary teams to work on the design and development of software. When dealing with AI systems, it is crucial to consider ethical aspects as well as success factors such as trust from the very beginning of the development process. As our research ought to indicate, there is an urgent need for further studies to examine the required competencies and viewpoints that should be included in order to ensure a human-centered artifact.

Furthermore, our research might indicate the urgent need of addressing challenges of AI and the findings of this research have to be made accessible to a broad public so as to ensure that a social discourse can take place. Humans have concerns about AI on different levels. Sometimes they have very specific objections, other times they have more of a negative feeling about AI. Researchers have certainly addressed some of the ethical issues of AI (van den Broek et al., 2019), but as our work here has shown, numerous guidelines and points of principle have yet to be discussed and defined for a thorough ethical, solution-oriented discourse. At present, there is a danger in that AI continues to be developed and disseminated without its specific problems and challenges having been specifically named or solved. This might point at the responsibility of organizations to play an active role in shaping AI design and development in ways that address user concerns but also account for ethical issues that most users will not even be aware of yet.

4.2 Future research

The following sections address the limitations of the six research papers presented above. They also offer first ideas on how to reach beyond these limitations and present starting points for future research.

4.2.1 Future research based on chapter 2: Behind the scenes of the Internet of Things

Several ideas for further research can be derived from the two articles about IoT. In section 2.1, the validation is based on 337 IoT devices, which resulted in five relevant categories used to validate IoT commerce affordances. It lies in the nature of emerging technologies, however, that their development is highly dynamic, so it has to be assumed that those five categories of IoT devices are limited and likely to change sooner rather than later. Future researchers may, therefore, indulge in new trends in IoT, add to our list of devices, and potentially extend our list of affordances. Furthermore, the examination is limited to B2C commerce, which leaves room for other researchers to answer our research question in the B2B context. Additionally, since IoT-commerce is a quite recent field of research, there are several starting points for further studies, for instance, differences between real and perceived affordances, customer acceptance of IoT-commerce, or negative side effects for customers in IoT-commerce.

Section 2.2 again underlines the manifold opportunities for further research in IoT. Since only scientific articles were included, leaving aside current discussions about ethical issues in practice, the scope of this research article is limited. Similarly reductive was the decision to neither questioning nor discussing whether or not there should be a separation between ethical and non-ethical issues of IoT. Instead, we relied on the assignment to ethics of the existing literature. In sum, four main areas of IoT ethics could be examined further. First, a thorough analysis of the extent to which known ethical issues of other technologies can be transferred to IoT ought to reveal the relevance of IoT-specific features and application contexts. Second, a deep dive into IoT features and application contexts could flag up further ethical issues that have not already been addressed in section 2.2. This includes an analysis of application contexts and features of IoT that are either currently emerging or will do so in the near future. Third, section 2.2 provides an overview of the known ethical issues of IoT, but to date, there is no in-depth analysis. Further research should focus particularly on the most relevant features causing ethical issues, namely *ubiquity, communication capability*, and *sensing and actuating capability*. Fourth, by taking a close look through the lens of opportunity, future researchers will see if and how IoT can contribute to solving ethical issues that are already familiar to us from other technologies.

In sum, it is crucial to recognize and promote the importance of further research on IoT, not only on the opportunities it affords us but especially on the challenges it poses. After all, due to its unique characteristics, the solutions that have been developed with other technologies in mind are only partially transferable. IoT provides manifold opportunities for individuals, and their exact dimensions will only become apparent in the future. To ensure that these opportunities can be exploited in full, a thorough examination of its potential for negative ramifications is required. There are various research opportunities including IoT. Given the rapid advances in technology, the most relevant research opportunities must be identified as a matter of urgency in order to guide further design and development of IoT in a direction that is both ethically sound and beneficial to humans. As mentioned above, despite technology convergence, our research might point fellow researchers in the direction of necessity and feasibility when dealing with IoT specifics.

Future research might establish links between individual IoT affordances and the related ethical issues. In a first step, based on the analysis of customer attitudes or expectations in a specific context, future research could reveal which affordances might be of special interest for organizations that wish to increase their market share. In a second step, each affordance could be examined for ethical issues that are particularly closely linked to the respective affordance. The results might guide users as well as organizations in making informed decisions about the usage and development of IoT devices under detailed consideration of the affordances and their related ethical issues. Perhaps the best starting point for this avenue of future research is the group of features that have been causing the greatest ethical issues, namely *ubiquity, communication capability*, and *sensing and actuating capability*.

Another starting point, albeit from a positive perspective, might be an examination of whether, and how, IoT features could help to solve or mitigate ethical issues of IoT. As our research has shown, IoT devices are varied and their disruptive potential has not yet fully unfolded (J.

P. Shim et al., 2020). This means that there is an opportunity or indeed a necessity for research on how IoT design and development can be driven not just by technological opportunities but also by human needs and the consideration of ethical implications.

4.2.2 Future research based on chapter 3: Behind the scenes of Artificial Intelligence

Certain promising research ideas can be derived from the four articles about AI. In section 3.1, the examination of concerns that individuals have about ADM is regionally limited, as all interviews were conducted in Germany. Hence, future research could explore a range of cultural comparisons that are bound to offer interesting revelations. Furthermore, our interviews were based on a selection of use cases, which limits the extent to which our results can be generalized. Further research could take this as a point of departure to examine, for instance, the differences between use cases or strive for more generalizability by detaching from the examples. Another idea worth exploring would be to examine our framework of concerns about the use of ADM with a view to finding correlations between individual concerns, for instance, by collecting and analyzing quantitative data.

Based on the ethical deep dive in section 3.2, the importance of further interdisciplinary research can be derived, such as examining what the limited agency attribution means in concrete terms as far as software development is concerned. Since we argue that the concept of moral agency should not be attributed to computational behavior, future researchers could dive deeper into conceptual differences between humans and computers. Our article indicates the first steps and direction this research might want to take, but a comprehensive concept has yet to be developed. Furthermore, this work should be supplemented with a detailed analysis of what a strict distinction between humans and computers would mean for human-computer interaction and the related ethical-normative structure should be analyzed in detail. The insights gained from such interdisciplinary research promise to contribute to the intense discussions about responsibility and accountability that are yet to be solved holistically.

Section 3.3 examines the role of domain expertise in XAI decision support systems, yet the validity of the results is somewhat limited by the uncontrollability of the surroundings in which the test subjects participated in our online experiment. Another limiting factor may be the sample size of 100 participants that might have influenced the results. Furthermore, our model could be expanded by including an analysis of explanation quality of the AI system explains its suggestions. Alternatively, it could be expanded by including additional variables from the theory of planned behavior, which is where parts of our model originated. It is worth noting that the chess context we chose for our study limits the generalizability of our research, yet future research could compensate for this by adding other contexts, such as healthcare or the judiciary, which are also at the center of discussions about AI systems. Moreover, the participants in such future studies ought to be screened for criteria other than their domain expertise so as to reveal the impact that, for instance, their character traits have on their trusting intention and behavior when they interact with XAI decision support systems.

The main limitation of the research paper in section 3.4 is the relatively small number of interviews. For future research, a significantly larger number of interviews would open up possibilities to compare different kinds of software development projects, such as the type of technology, the project size, or the team size. Based on those insights, conclusions could be drawn about potential advantageous characteristics of software projects and their environmental conditions that favor ethical development. Furthermore, our ethical software development framework could be evaluated in greater detail, for instance in a range of experiments, which would allow further refinement of the framework.

To sum up, future research could focus on specific use cases and contexts relevant to AI, such as high-impact and high-risk tasks. Alternatively, it could strive for generalizability by significantly expanding qualitative data collection and, where expedient, extend its methodology to quantitative data collection and analysis. There is an urgent need for more interdisciplinary research with ethicists, but also with other disciplines such as law, in order to truly advance research and practice that is required to deal with the dark-side aspects of AI.

Our research might point at the necessity of further analyzing the role of AI stakeholders in its design and development. As one interviewee in section 3.4 said, it is not easy to neglect offers of ethically questionable projects, as there is a high probability that others will accept the offer should he decline to do so. In accepting such offers, he at least ensures that he can attempt to influence a questionable project in a positive sense. As our results might suggest there has to be a detailed examination of the responsibilities of all stakeholders, for instance, organizations, employees, and users, but also politicians, in order to safeguard a joint venture of ethically sound consideration and conduct in AI design and development. Further research could potentially examine the specific roles of stakeholders as well as analyze their impact, for instance in experimental settings.

A final point worth making here is that knowledge about the user of an AI system would appear to be of the utmost importance in the design and development phase so as to improve the odds of a successful market launch. Another key factor is user trust (Yan et al., 2011), as numerous concerns about AI systems are accompanied by the concrete need to explain AIs suggestions to its users. To build trust, organizations need to know the characteristics of the target group. Only then can they successfully address potential concerns and provide customized explanations (Cooper, 1999). Future researchers might, therefore, conduct a detailed characterization of user groups of AI systems and analyze their specific concerns alongside their knowledge in selected contexts. This would provide reference points for organizations as they attempt to design AI systems that account for the needs of humans.

4.3 Conclusion

In summary, this dissertation contributes to knowledge development for IS about opportunities, but with particular regard to the challenges that IoT and AI pose along the socio-technical continuum. With its focus on those challenges, this dissertation enriches the meaningful and increasingly important research field on the dark side of IS, which has recently received somewhat more attention but remains eclipsed by research on opportunities. Through expectations about technologies and users' behavior, individuals' influence on the direction of digitalization increases. To account for this influential role of users, this dissertation looks at the opportunities and challenges from the perspective of individuals. By situating six research articles along the socio-technical continuum proposed by Sarker et al. (2019), this dissertation provides insights into IoT's affordances for customers, examines the challenges for IoT and AI (particularly those of an ethical nature), and offeres solution approaches for the key concept of trust in AI systems and for the inclusion of ethics in software development. In doing so, this dissertation follows calls for IS research along the socio-technical continuum that should serve the needs of humanity (Majchrzak et al., 2016; Sarker et al., 2019). IS research should assume the role of translator between the social and the technical ends of the spectrum. Otherwise, we face the danger of unclear responsibilities, especially when researching the challenges of emerging technologies and their solution approaches (Franke & Zoubir, 2020). It is crucial to avoid being overwhelmed by the sheer magnitude and rapid pace of the digital transformation but actively take on the role as shapers of emerging technologies, starting by defining needs rather than "chasing possibilities" (Franke & Zoubir, 2020). Hopefully, this dissertation will provide valuable theoretical and practical contributions to further, human-needs centered research on emerging technologies along the socio-technical continuum.

References

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, *21*(4), 12–17.

Bayer, S., Gimpel, H., & Rau, D. (2021). Iot-commerce - opportunities for customers through an affordance lens. *Electronic Markets*, *31*(1), 27–50.

Cooper, A. (1999). *The inmates are running the asylum*. Sams.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences*, *374*(2083).

Floridi, L., & Sanders, J. W. (2001). Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*, *3*, 55–66.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*, 349–379.

Franke, T., & Zoubir, M. (2020). Technology for the people? Humanity as a compass for the digital transformation. *Wirtschaftsdienst*, *100*(13), 4–11.

Gregor, S., & Hevner, A. (2013). Positioning and presenting design science research for maximum impact. *Management Information Systems Quaterly*, 337–355.

Jeong, S., Kim, J.-C., & Choi, J. Y. (2015). Technology convergence: What developmental stage are we in? *Scientometrics*, *104*(3), 841–871.

Johnson, D. G., & Verdicchio, M. (2018). Ai, agency and responsibility: The vw fraud case and beyond. *AI & SOCIETY*, *34*(3), 639–647.

Karwatzki, S., Trenz, M., Tuunainen, V. K., & Veit, D. (2017). Adverse consequences of access to individuals' information: An analysis of perceptions and the scope of organisational influence. *European Journal of Information Systems*, *26*(6), 688–715. https://doi.org/10.1057/s41303-017-0064-z

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. *Management Information Systems Quarterly*, *40*(2), 267-277.

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the is discipline: Its historical legacy and its continued relevance. *Management Information Systems Quarterly*, *43*(3), 695–719.

Shim, J. P., Sharda, R., French, A. M., Syler, R. A., & Patten, K. P. (2020). The internet of things: Multi-faceted research perspectives. *Communications of the Association for Information Systems*, *46*, 511–536.

Stahl, B. C., & Rogerson, S. (2009). Landscapes of emerging ict applications in europe. *Proceedings of the Eighth International Conference of Computer Ethics: Philosophical Enquiry*.

van den Broek, E., Sergeeva, A., & Huysman, M. (2019). Hiring algorithms: An ethnography of fairness in practice. In *Proceedings of the 40th International Conference on Information Systems (ICIS),* Munich, Germany.

Verbeek, P.-P. (2006). Materializing morality. Design ethics and technological mediation. *Science, Technology, & Human Values*, *31*, 361–380.

Verbeek, P.-P. (2011). Moralizing technology. Understanding and designing the morality of things. *Chicago: Univ. Of Chicago Press*.

Verbeek, P.-P. (2014). Some misunderstandings about the moral significance of technology. *The Moral Status of Technical Artefacts, Edited by Peter Kroes and Peter-Paul Verbeek. Dordrecht: Springer*, 75–88.

Verbeek, P.-P. (2017). Designing the morality of things: The ethics of behaviour-guiding technology. *Designing in Ethics, Edited by Jeroen Van Den Hoven, Seumas Miller and Thomas Pogge. New York: Cambridge Univ. Press*, 78–94.

Yan, Z., Kantola, R., & Zhang, P. (2011). A research model for human-computer trust interaction. In *10th ieee international conference on trust, security and privacy in computing and communications*, Changsha, China.