Task delegation from AI to humans: A principal-agent perspective

Conference Paper · December 2023

CITATIONS
4

READS
2,346

5 authors, including:

Tobias Guggenberger
University of Bayreuth
42 PUBLICATIONS **567** CITATIONS
SEE PROFILE

Luis Lämmermann
University of Bayreuth
9 PUBLICATIONS **77** CITATIONS
SEE PROFILE

Nils Urbach
Frankfurt University of Applied Sciences
272 PUBLICATIONS **8,735** CITATIONS
SEE PROFILE

Anna Walter
University of Bayreuth
1 PUBLICATION **4** CITATIONS
SEE PROFILE

Task delegation from AI to humans: A principal-agent perspective

Completed Research Paper

Tobias Guggenberger
University of Bayreuth,
Fraunhofer FIT,
Bayreuth, Germany
tobias.guggenberger@fim-rc.de

Luis Lämmermann
University of Bayreuth,
Bayreuth, Germany
luis.laemmermann@uni-bayreuth.de

Nils Urbach
Frankfurt University of Applied Sciences, Frankfurt, Germany
nils.urbach@fb3.fra-uas.de

Anna Walter
University of Bayreuth,
Bayreuth, Germany
anna.walter@magenta.de

Peter Hofmann
FIM Research Center,
Bayreuth, Germany
peter.hofmann@fim-rc.de

Abstract

Increasingly intelligent AI artifacts in human-AI systems perform tasks more autonomously as entities that guide human actions, even changing the direction of task delegation between humans and AI. It has been shown that human-AI systems achieve better results when the AI artifact takes the leading role and delegates tasks to a human rather than the other way around. This study presents phenomena, conflicts, and challenges that arise in this process, explored through the theoretical lens of principal-agent theory (PAT). The findings are derived from a systematic literature review and an exploratory interview study and are placed in the context of existing constructs of PAT. Furthermore, this article paper identifies new causes of tensions that arise specifically in AI-to-human delegation and calls for special mechanisms beyond the classical solutions of PAT. The paper thus contributes to the understanding of autonomous AI and its implications for human-AI delegation.

Keywords: Delegation, artificial intelligence, human-AI collaboration, principal-agent theory

Introduction

Technological advances in recent decades have led to ever-improving human-AI collaboration and automating tasks (Peeters et al. 2021). Human-AI systems today exceed the individual capabilities, advantages, and disadvantages of both humans and AI (Dellermann et al. 2019; Jarrahi 2018). Increasingly autonomous AI artifacts with superior quantitative, computational, and analytical capabilities (Jarrahi 2018) are no longer limited to performing tasks on humans' behalf. Instead, they are performing tasks more and more autonomously as entities that guide human actions and are even changing the task delegation direction between humans and AI artifacts (Baird and Maruping 2021; Harms and Han 2019; Wesche and Sonderegger 2019). This shift in delegation direction leads to a new distribution of tasks and roles within human-AI systems, for instance, expressing a change in task leadership of AI artifacts over humans. As the

research indicates, human-AI systems can achieve better results when AI artifacts take the lead role and delegate tasks to a person rather than the other way around (Fügener et al. 2022). While AI artifacts follow an efficient delegation strategy, humans tend to make worse delegation decisions since they cannot assess their capabilities regarding specific tasks' difficulty. For this reason, AI artifacts seem better suited for delegation ownership in human-AI systems. In conjunction with their capabilities, which are also becoming more vital in other respects, AI artifacts are acquiring a certain amount of task and process ownership (Ågerfalk 2020; Fügener et al. 2022; Harms and Han 2019).

The increasing autonomy shift in the decision-making of delegation toward AI artifacts is being recognized in multiple application areas. For instance, the breast cancer screening application of the company Vara[1] has demonstrated how increasing AI autonomy is changing task delegation between physicians and AI artifacts during medical decision-making processes. In the concrete application example, the AI artifact decides whether the physician should evaluate a medical case (i.e., mammogram). While algorithmic assessments with high certainty were executed automatically, only those with lower confidence were referred to the radiologist. The AI-led triaging process resulted in higher efficiency and lower workload for the radiologist while maintaining higher sensitivity and specificity than either an AI or a radiologist working alone (Leibig et al. 2022). In addition, autonomous AI artifacts with delegation ownership are used in decision-making in the financial sector, such as in software that automatically reallocates assets or acts as an underwriter that can process loans (Berente et al. 2021; Lee and Shin 2018; Markus 2017).

This new phenomenon of AI artifacts making delegation decisions and transferring tasks to humans has significant consequences for human-AI collaboration. For instance, it fosters uncertainty within organizations since it dramatically alters the previous structures and nature of human-AI collaboration and leads to fundamental changes in organizational design and coordination (Benbya et al. 2020; Wesche and Sonderegger 2019). The uncertainty manifests itself for instance in unclear process control and accountability, as humans lose control of the process and decision-making when transferring the delegation ownership to an AI agent (Fügener et al. 2022; Steffel et al. 2016). Further, the opacity of AI decisions and underlying rules can lead to a lack of trust and undesirable human behaviors that oppose AI (Vössing et al. 2022). These new interaction dynamics and conflicts between humans and AI artifacts as agentic entities working together must be understood if one is to develop human-AI systems led by an AI artifact, considering both organizational and human subordinates' interests (Baird and Maruping 2021; Wesche and Sonderegger 2019).

While human-to-AI delegation has been recognized as a crucial research topic, AI-to-human delegation is still unexplored. The research lacks strong theorizing concerning task delegation from digital principals to human agents (Baird and Maruping 2021) and implications for organizations on how to deal with the potentials of increased automation in AI-to-human delegation. How AI systems with delegation ownership must be designed to augment humans and facilitate human-AI delegation is poorly understood (Rai et al. 2019; Vössing et al. 2022). We consider explicit research in this area to be highly relevant, because human-AI collaboration with increasingly autonomous AI artifacts will become more dominant in the future owing to its superior results (e.g., Dellermann et al. 2019; Hemmer et al. 2021; Vössing et al. 2022; Xu et al. 2023). Our goal here is to contribute to the understanding of how human-AI collaboration in delegation processes changes when AI takes over the ownership for delegation, reducing the uncertainties regarding such collaboration faced by companies. We ask:

What are the key tensions that arise from delegating tasks from artificial intelligence to humans, and what factors contribute to these tensions?

We explore this topic through principal-agent theory (PAT) as our theoretical lens. This theory is used to describe relationships as contracts where *"one or more persons (the principal(s)) engage another person (the agent) to perform some service on their behalf which involves delegating some decision making authority to the agent*" (Jensen and Meckling 1976, p. 5). It suits our study purpose well because it shows the nature of cooperation and conflict between the human and AI artifacts as agentic entities, as both collaborate to achieve objectives (Baird and Maruping 2021; Emirbayer and Mische 1998). To date, the research has applied this theory predominantly to the relationship between humans as principals and algorithms working for them as agents (e.g., Kim 2020, Borch 2022). We study the paradigm shift toward

[1] https://www.vara.ai/

AI artifacts becoming the principal, delegating tasks to humans as agents, through the lens of PAT. We conducted a systematic literature review and an exploratory interview study to examine AI-to-human delegation from a principal-agent perspective, highlighting the arising phenomena, conflicts, and challenges. By extending the principal-agent view beyond consideration of only human principals and examining the resulting phenomena, we contribute to the literature on agency theory. Furthermore, applying an agency theory perspective to human-AI delegation will increase our understanding of AI, thereby contributing to the emerging literature on human-AI interaction.

Theoretical Background

AI, humans, and delegation

In human-AI delegation, many terms are used ambiguously. We clarify our understanding as follows. First, we refer to AI artifacts as systems that *"perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem-solving, decision-making, and even demonstrating creativity"* (Rai et al. 2019, p. iii). This includes abilities that were limited to humans, such as communication using natural language, forming abstractions, and improving itself (Bawack et al. 2019). AI has led to a paradigm shift in the relationship between humans and machines, since systems are no longer based solely on fixed logical rules and no longer only respond to human inputs (Farooq and Grudin 2016; Xu et al. 2023). The research emphasizes that, although machines can outperform humans far more often than vice versa, the future of work will be humans and AI working together synergetically (Jarrahi 2018). Settings where AI artifacts are used to augment rather than replace human contributions have been shown to lead to the most significant performance improvements in organizations. At the same time, the human remains at the center of considerations, and human-AI collaboration should be focused on them (Fügener et al. 2022; Hemmer et al. 2021; Jarrahi 2018).

Delegation is a key feature that enables efficient work and provides productivity, flexibility, and job satisfaction (Griffiths 2005; Gur and Bjørnskov 2017), because one party alone typically does not have the resources, capabilities, or knowledge to achieve their goals on their own (Griffiths 2005). We define delegation as *"the transfer of rights and responsibilities for task execution and outcomes to another"* (Baird and Maruping 2021, p. 317). Used appropriately, delegation can improve the decision speed and quality, reduce managerial overload, enrich a subordinate's work and intrinsic motivation, and provide opportunities for leadership development by a subordinate (Yukl and Fu 1999).

The research has mainly focused on human-to-AI delegation, with people being the leading entity to which the AI artifact is subordinate (Fügener et al. 2022). However, AI artifacts' ability to assess their own capabilities and those of a human partner, and to recognize complementarities between them, have created new opportunities in task delegation (Fügener et al. 2022; Goldbach et al. 2019). Studies that have expanded the view of AI artifacts as digital agents and that recognize AI artifacts as digital principals in human-AI systems include Fügener et al. (2022), Ågerfalk (2020), and Baird and Maruping (2021). These studies state that the digital entity can also be the principal, not just the agent, and demonstrate this paradigm shift's benefits and relevance. Regarding delegation between humans and nonhuman entities, Baird and Maruping (2021) state that human and information systems can be the delegating units. However, Fügener et al. (2022) showed that AI artifacts perform better in delegating tasks to humans than vice versa. Humans are imperfect at judging their capabilities and task difficulty, which leads to lousy delegation decisions. The performance of AI artifacts that delegated tasks to humans in situations of uncertainty was higher than when these artifacts operated alone, rendering this type of collaboration economically attractive. It improves the workplace and human motivation, as employees can spend more time on challenging tasks and less time on mundane ones. In this setting, an AI artifact would be a human's assistant that removes distractions from the actual work (Fügener et al. 2022). We refer to this direction of delegation as AI-to-human delegation.

Following Baird and Maruping (2021), we consider the human as the individual who interacts with the AI artifact to accomplish a task, rather than referring to its programmer who is upstream in the creation of the artifact. Regarding the delegation object, a distinction can be made between the delegation of either cognitive tasks, the execution of subordinated tasks, or both to the AI artifact (Loi and Spielkamp 2021). We focus on the second type of delegation mentioned by Loi and Spielkamp (2021) - the delegation of the execution of subordinated tasks within human-AI systems. Thus, we consider delegation between AI

artifacts and humans, whose delegated tasks are executable by both entities. This also distinguishes our study from studies in the algorithmic management research stream as we do not limit the role of the AI artifact to managerial coordination functions but broaden this view by following the recent emergence of conceptualizing AI artifacts as autonomously acting, interacting, learning, and adapting to their environment. AI artifacts can now make their own decisions without much human intervention and perform tasks without each step being procedurally defined beforehand (Ågerfalk 2020; Baird and Maruping 2021; Dattathrani and De' 2023; Schuetz and Venkatesh 2020).

Principal-agent theory

The primary theoretical lens we use to explore human-AI delegation is PAT. It suits our work well, since it deals with cooperation and conflicts between agentic entities working together (Baird and Maruping 2021), and remains valid and applicable in contexts with nonhuman agents (cf. Baird and Maruping 2021; Borch 2022; Chen and Barnes 2014; Kim 2020; Lazányi 2018; Wesche and Sonderegger 2019). PAT examines the relationship between at least two contracting parties (i.e., principal and agent) in conditions of unequally distributed information and divergent goals, considering the uncertainty and risk propensities of the participants (Bergen et al. 1992; Jensen and Meckling 1976). In these relationships, the principal delegates decision-making authority to an agent, who performs services on their behalf (Jensen and Meckling 1976). It focuses on determining the most efficient contract to govern this relationship and the associated conflicts caused by the division of labor and responsibilities between the principal and the agent (Bergen et al. 1992; Schneider 1987).

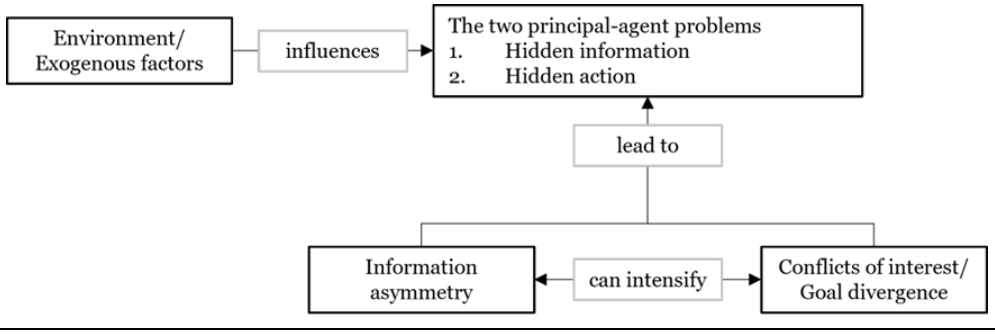

Figure 1. The principal-agent problems and their causes

At the center of PAT lies two principal-agent problems (PAPs), hidden information and hidden action, also referred to as adverse selection and moral hazards (Arrow 1986). Hidden information arises pre-contractually and refers to the principal's difficulty in determining whether an agent has the qualities and skills necessary to perform a task in their interest (Bergen et al. 1992; Fayezi et al. 2012). Hidden action emerges post-contractually and describes the agent's activities that may be in their own but not the principal's interest (Bergen et al. 1992). As shown in Figure 1, both problems arise owing to information asymmetry and conflicts of interest between principal and agent, and can be influenced by the environment and exogenous factors (Bergen et al. 1992; Keil 2005; Peterson 1993).In agency theory, information is seen as a commodity that can be exchanged (Eisenhardt 1989). An agent has superior knowledge and is often reluctant to share information with the principal (Bergen et al. 1992; Schneider 1987). When the interests between the principal and the agent do not align, and there is a goal mismatch, the agent may hide information from the principal, which causes information asymmetry and increases potentials for opportunistic behaviors. The reasons agents often are unlikely to behave in their principal's best interest lie in PAT's assumptions about the entities' behaviors: they are self-interested, and have different risk preferences and bounded rationality (Heaslip and Kovács 2018; Tate et al. 2010). In this context, it can be assumed that both parties to the relationship try to maximize their benefits, leading to a divergence between the agent's decisions and the principal's interests (Jensen and Meckling 1976).

The most appropriate solution to PAPs depends on the two parties' goals and risk preferences, the types of delegated tasks, and the level of environmental uncertainty (Bergen et al. 1992). However, the efficient management of agency problems is imperative to any principal-agent relationship (PAR) (Heaslip and Kovács 2018), and the literature has agreed on generally good solutions to address the issues: The *hidden information* problem can be addressed by gathering information about the agent through screening

activities, examining signals from agents, or increased information sharing between the two parties (Bergen et al. 1992; Fayezi et al. 2012; Pavlou et al. 2007; Poth and Selck 2009). The *hidden action* problem can be reduced by signals, bonding, and the monitoring of behavior or performance (Pavlou et al. 2007). Further, the alignment of incentives through for instance outcomes-based rewards is emphasized as an essential way to reduce hidden actions, since it increases goal congruence (Fayezi et al. 2012; Heaslip and Kovács 2018; Morris et al. 2020).

Although the PAPs were initially formulated to study the separation of ownership and control (Berle and Means 1932), it is ubiquitous and applicable to a broader range of applications beyond employment relations, including cooperation and delegation (Borch 2022; Pavlou et al. 2007). Thus, PAT has been utilized by various disciplines, each with different interpretations of and solutions to the PAPs (Kim 2020; Shapiro 2005). Approaches to PAT have long focused only on human-centered relationships, with nonhuman agents (e.g., computer algorithms) left out of consideration (Kim 2020). Yet, owing to their superior quantitative, computational, and analytical capabilities (Jarrahi 2018), researchers have begun to also recognize nonhuman agents (e.g., intelligent computer systems) as agents within PAT (cf. Baird and Maruping 2021; Chen and Barnes 2014; Wesche and Sonderegger 2019). Accordingly, researchers have applied PAPs to human-to-AI delegation, with AI systems acting as agents of human principals, for instance in automated trading (Borch 2022), algorithmic governance (Kim 2020), and autonomous driving (Lazányi 2018).

As autonomous systems increasingly issue instructions to humans as principals, new relationships and interactions are emerging (Jennings et al. 2014). So far, there has been little to no research about these relationships between digital, intelligent principals and human agents and their implementation into organizations. However, what has already been shown is that information asymmetries exist between the AI artifact as principal and the human agent due to their different natures, information knowledge, and capabilities (Baird and Maruping 2021; Vincent 2021). Furthermore, principal-agent roles and the balance of power shift as humans lose sovereignty over the process and decision-making to a technological entity (Fügener et al. 2022), which can lead to conflicts of interest (Fügener et al. 2022; Vössing et al. 2022). Yet, it remains uncertain what this collaboration will look like when implemented in organizations, how the PAPs in AI-to-human delegation are changing and dealt with, and which new conflicts arise (Benbya et al. 2020; Wesche and Sonderegger 2019).

Research Method

Our methodological approach was twofold: To integrate insights from theory and practice and address our research objective, we combined a systematic literature review (SLR) guided by Webster and Watson (2002) with qualitative semi-structured interviews (Myers and Newman 2007). To get an overview of the current state of research in the field of delegating relationships between humans and AI artifacts, we first synthesized existing literature on the topic. Thus, we used the following search string, consisting of two search terms: ([All Fields](“Artificial Intelligence” OR AI OR “collective intelligence” OR “hybrid intelligence” OR “human-machine” OR “human-computer” OR “human-IS”) AND [Abstract](Delegat* OR “decision-making structures”)). With our first search term, we defined our technical scope around AI and the human-AI construct. The keywords in the second search term and narrowing down the search to the abstracts in the literature ensured that the focus was on delegation. We searched the databases the Association for Information Systems eLibrary (AISeL) for the information systems perspective, the Web of Science (WoS) database for the broader scope, and the ACM Digital Library for the computer science perspective. The initial search resulted in 970 papers. After removing duplicates (-25), title- (-811), abstract- (-75) and full-text screening (-46), as well as forward and backward search (+6), we identified 19 relevant papers. In our review, we excluded publications where human and artificial intelligence entities did not exhibit characteristics consistent with a PAR or those that lacked an exploration of delegation or decision-making mechanisms, dimensions, or conflicts. While the research has predominantly been concerned with human-to-AI delegation, considering AI only as a human principal’s agent, we were able to explore existing knowledge on delegation structures, mechanisms, and factors that affect and determine the design delegation between humans and AI artifacts from the relevant papers in the SLR.

We then built on the identified state of the research by performing an in-depth interview study to expand our theoretical understanding of the phenomenon of AI artifacts being in the principal role and delegating ownership. The SLR results first served as a starting point and as interpretive devices for our qualitative

study, following the research approach of sensitizing concepts (Bouwen 2006; Glaser 1978; Padgett 2003). These concepts are background thoughts that determine the general research problem (Charmaz 2000), and they involve the researchers' attempts to discover, understand, and interpret the happenings in a research context. Thus, they serve as a foundation or guideline for analyzing research data, for instance for developing thematic categories from the data (Bouwen 2006).

Semi-structured interviews with experts enable one to focus on a topic and provide participants with in-depth information, while allowing them to reflect on their experiences and perceptions of a case (Myers and Newman 2007). As interviews are specifically appropriate for understanding experiences, opinions, attitudes, values, and processes (Rowley 2012), we found them a suitable method to validate and extend the outcome of our SLR and explore the topic in greater depth. Following Bogner and Menz (2009, p. 55), we use *expert* to refer to a person with a "*specific configuration of knowledge*". We considered two types of experts. First, we looked for people who understand autonomous systems in practice and, in the best case, have been involved in implementing agent systems where AI artifacts have some autonomy and task responsibility. Regarding the tasks in this setting, we sought to gather various perspectives by interviewing for instance software developers, product owners, and process owners, following the approach of Eisenhardt and Graebner (2007), who suggest using various highly knowledgeable informants with multiple angles on the subject. We did online research to select potential experts with practical experience on this subject and identify companies that are actively using or planning to use AI artifacts in task ownership. Second, since we aim to investigate how the human-AI PAR manifests when the artifact is in the principal role, we also sought experts in the PAT field, explicitly looking for people who have dealt with digital, nonhuman agents in the context of PAT. We screened the articles published on this topic and noted the authors as potential interview partners. Owing to the specificity of the required expert knowledge, we also used the snowballing method for data collection. All interviewees were asked whether they could name others with the required knowledge and experience (Goodman 1961; Noy 2008). Table 1 shows the experts in our interview study.

Experts	**Professional titles**	**Field of expertise**	**Type of organization**	**Duration (min.)**
E1	Co-Founder and CDO	Data science	Healthcare startup	46
E2	Professor, economic sociology and social theory	Technologies in financial markets, collective behavior	University	37
E3	Professor, human factors, industrial and organizational psychology	Algorithmic management, technology design, automation, and work	University, startup	53
E4	Co-Founder and CEO	Business development	Healthcare startup	46
E5	IT Manager, business and digital solutions	Automation of business processes	Global manufacturer of medical products	61
E6	Co-Founder and CTO	AI, data, and analytics	Hybrid AI startup	44
E7	Data scientist	Data science	Company in the petroleum sector	37
E8	Managing consultant, data science	Human-AI interaction	IT and consulting firm	46
E9	Co-Founder and CEO	AI, human factors, and quantitative methods	Collective intelligence startup	35
E10	Researcher	Algorithmic management, human-AI delegation	Research institute	42*
E11	Researcher	AI literacy, human-AI delegation, algorithmic management	Research institute	42*
E12	Researcher	AI management, technology convergence	Research institute	34
E13	Postdoc in social, organizational, and economic psychology	Interactions and leadership in digital work	University	32

Table 1. Overview of the interviewed experts * group interview

We conducted the semi-structured interviews based on a predefined interview guide we had developed iteratively. We derived the overarching theme blocks in the guide from the research objective, the research question, and the SLR's results. After establishing a shared understanding of PAT and the terms used, we

interviewed the experts with experience in practical design, communication and information, and human behaviors and responses in implementing AI in delegation. We asked theory experts how the two primary constructs of PAT – information asymmetry and conflicts of interest – change when the principal is no longer a human but an AI artifact. In the last step, the interviewees were allowed to mention points that had not yet been addressed but were considered necessary in AI-to-human delegation. We recorded and transcribed all interviews with our interviewees' consent.

For subsequent data analysis, we followed the systematic approach by Gioia et al. (2013), which seeks to ensure qualitative rigor in inductive research and is often used when a deeper understanding of organizational processes and dynamics is required. It is well suited to our research goal because it allows unstructured qualitative datasets to be processed, relevant categories and relationships between them to be formed, and new concepts, ideas, and theories to emerge (Corbin and Strauss 2015; Gioia et al. 2013). We adopted the iterative three-step coding process to study the same phenomenon of interest at different abstraction levels (Gioia et al. 2013). In phase 1, the transcripts were screened through open coding, and relevant information for our research question was highlighted to understand the data's breadth and depth. Through multiple iterations, 119 descriptive codes were identified, and 24 first-order concepts were assigned, maintaining the informant-centric terms' integrity. This set was modified during the data analysis to reflect new information and exclude ideas that did not seem relevant to our studies. In phase 2, when moving from open to axial coding, the existing open codes and categories were related, and relationships between them were established (Corbin and Strauss 1990; Gioia et al. 2013). The broad first-order concepts from the underlying data were grouped under more abstract and theoretical themes, resulting in 10 second-order themes. In phase 3 (selective coding), we transformed our second-order themes into four overlying aggregate dimensions by examining how the axial codes fit together (Corbin and Strauss 2015). PAT and its concepts provided the structure for our aggregate dimensions here, as we wanted to investigate the human-AI PAR in AI-to-human delegation. Thus, this step involved moving back and forth between second-order themes, the SLR's findings, and concepts from PAT to identify topics that were inadequately represented in the literature yet. We also conducted three collaborative coding workshops so as to increase the reliability and validity (internal and external) of the categories and constructs derived from the dataset (Lombard et al. 2002). We provide our coding system as supplementary material in Guggenberger et al. (2023).

Results

Based on the SLR and the interview study, we identified four dimensions that conceptualize the PAR in AI-to-human delegation, subsuming our findings along the three classical causes and influences of PAPs, which are complemented by a fourth construct specific to AI-to-human delegation. As shown in Figure 2, our results shed light on how different *information asymmetries* between AI artifacts and humans impact on the delegation decision and place specific demands on the delegation relationship. Also, we depict which *conflicts of interest* types arise between AI artifacts and humans. In the third dimension, we look at the influences of the *environment and exogenous factors* on the delegation relationship. While the first three dimensions are the classic concepts within PAT, we identified the fourth dimension, *human attitude toward the AI artifact as principal*, as a new cause of PAPs specific to AI-to-human delegation. Further, we observe various new phenomena within the different dimensions that are unique in AI-to-human delegation and lead to PAPs there.

Information asymmetries

One relevant information asymmetry between humans and AI artifacts in the literature is the lack of transparency of AI artifacts (Araujo et al. 2020; Candrian and Scherer 2022). The rules for delegation are set within the AI's implicitly defined decision model, which is inscrutable to a human agent (Vössing et al. 2022). In many application areas, people are more willing to follow instructions if they understand the background and goals that are being pursued with them (Baird and Maruping 2021; Pasquale 2015). The lack of transparency about the principal's intentions and procedures is much more prevalent in AI-to-human delegation than in purely human relationships, *"since people can at least project ideas from one person to another and can get a sense of another person's thinking"* (E2). Further, our dataset shows that a critical issue in AI-to-human delegation, which arises from information asymmetries and differs from purely human delegation relationships, is the balance of control between a nonhuman principal and a human agent. Although AI artifacts do not yet have legal autonomy, owing to regulations, they are already

used in settings where humans follow their instructions and recommendations as subordinates (E2; 3; 6; 8; 10).

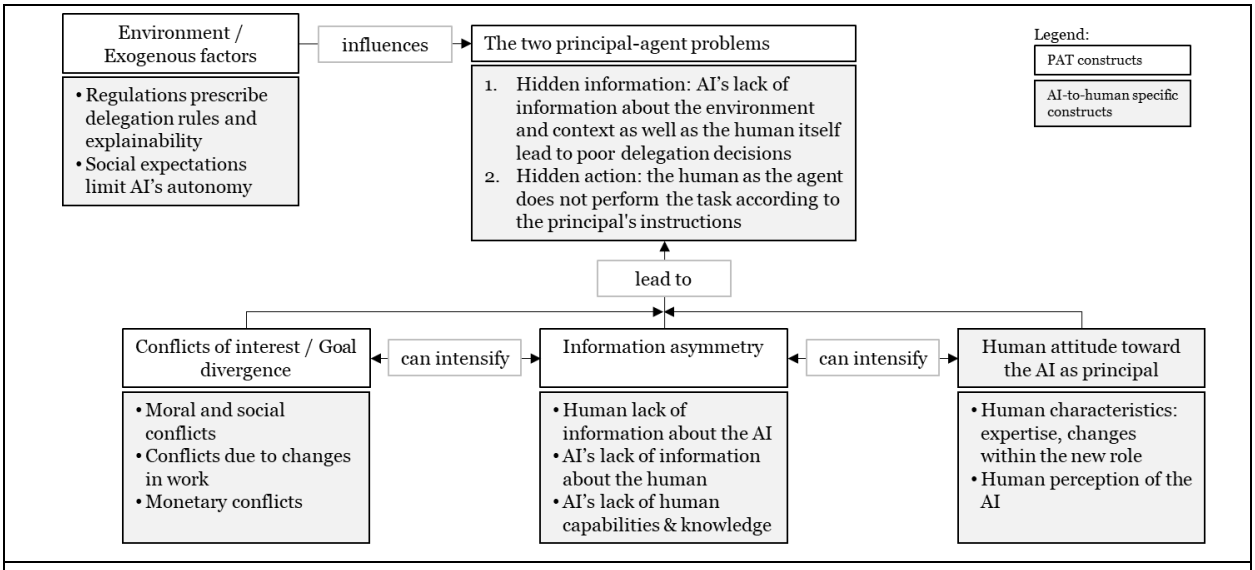

Figure 2. The principal-agent problems in AI-to-human delegation

In line with the findings in the literature (Fuchs et al. 2022; Parry et al. 2016; Wilson and Daughtery 2018), our results show that, even if an AI artifact is in an autonomous leadership role and is superior to humans, there must be the ability to relinquish control to a human-in-the-loop (E1-13). The difficulty is determining when the artifact needs to return power and control to humans and the mechanisms through which this can be ensured (Jennings et al. 2014; E1; 2; 4). The question arises which safeguards are required for humans and AI artifacts when the artifact is given delegation ownership yet lacks certain information (Jennings et al. 2014; E2; 3; 7; 9); it relates to the tension between autonomous AI artifacts with delegation ownership that give humans tasks and instructions, but over which humans must still have control in the end.

"Specifically, we had to build in mechanisms that would warn the human when the artifact was no longer performing well, that is, when either the predictions it made were bad, or some other error had occurred in the software or hardware. So that the human can then no longer rely on the artifact." E1

While E9 states that humans must know these systems' limits, E6, 8, and 11 emphasize that this is not always possible with self-learning systems, which are based on rules set by humans but occasionally also develop their own ones that are not always transparent and comprehensible to humans (E1-13). Owing to their opacity, people often have little to no insight into the systems, the rules underlying their decisions, and whether or not they act within their requirements' scope. This was also confirmed by E1, 2, 3, 4, 10, and 13 which state that people still have control because they can decide whether or not to follow the AI artifact's instructions. However, this opacity makes it hard to assess when they should explicitly not do so. Although explainable AI is a vast and crucial research area, satisfactory solutions that provide understandable decision rationales to explain multidimensional AI decision models' outputs remain limited (Benbya et al. 2020; E2; 3; 6; 7; 8; 9). E6 and 8 emphasize that explanations are only approximations and can be manipulated, distorted, and biased. The fact that this principal-agent information asymmetry cannot yet be reduced by explainability increases the need for mechanisms to ensure appropriate actions when an AI artifact is the principal (E1; 2; 4; 6; 9). This becomes particularly relevant when AI, as a black box, risks violating regulatory requirements or endangering human safety, for instance when autonomous AI artifact with delegation ownership is used in financial trading or cancer detection (E1; 2; 4; 8).

Another asymmetry identified in the literature relates to information that humans have but AI artifacts do not (Abbasi et al. 2022; Beer et al. 2014; Shrestha et al. 2019; Vincent 2021). The principal is now an algorithm that exceeds humans in several complex analytical tasks using computational methods and big data; at the same time, it does not yet have the necessary information and capabilities for many uncertain, ambiguous ones (Vincent 2021). In contrast, humans have situational awareness and intuition and can use

their involved, embodied, and reactive experiences to solve nonroutine problems (McClure 2014). Further, the AI artifact cannot yet understand and respond to emotional cues in ways humans can, which is why it cannot respond to them with the same emotional intelligence as a human superior (Abbasi et al. 2022; E3; 6; 8; 12; 13). Moreover, regarding specific application areas, an AI artifact cannot yet put data into the appropriate context (Hemmer et al. 2022; E4; 5; 7). It acts based on an optimization function to achieve a specific end goal without sufficiently considering the broader context of action, which is relevant in some use cases (E4; 6; 13). E5 and 6 state that, owing to this lack of contextual awareness, specific control mechanisms are needed before a human agent executes a task, for instance, plausibility checks or setting constraints that ensure that an AI artifact's performance is within a specific range. Also, building on the SLR's results, our study shows that AI artifacts often lack certain situation-specific information, which is required to perform tasks, and the ability to decide who should perform a task (Jennings et al. 2014; Vincent 2021; E8). Humans often have access to additional data not made available to an artifact, since not all data may be digitized and provided for training, for technical or economic reasons (Hemmer et al. 2022). E1-7 identify this information asymmetry as problematic. Still, it was also pointed out that AI artifacts can overcome this in some situations by instructing humans, who are the agents that interface between the digital and physical worlds, as data providers (Jennings et al. 2014; E1; 4; 5; 8).

"So, if you notice that something doesn't fit with the data or that we're missing additional information, then the human is instrumentalized to provide this additional information." E4

The prerequisite is that the AI artifact becomes aware of which data is missing and when (E4). E1 and 4 also emphasize that the autonomy of AI as a principal requires recognizing its potentials and need for improvement. Through human data feedback to the artifact, it can actively improve its performance and can learn to identify anomalies and patterns in datasets. This both improves its performance and expands its information basis for delegation decisions (E6; 7). This can also compensate for information asymmetries that arise when AI artifacts are used in a dynamic environment. When external factors change, an artifact must be adapted to these changes through new data inputs (E2; 4; 7; 9; 12). But this requires specific interactivity between the two entities (i.e., an information flow is possible) and, more importantly, the AI artifact must be able to process, utilize, and respond to the input (E6). However, in active learning, humans influence an AI artifact's performance and behaviors through feedback. These potential opportunistic behaviors by humans owing to their conflicts of interest must therefore be considered when designing human-to-AI feedback information flows (E9).

To date, the literature has almost exclusively highlighted information asymmetries between humans and artifacts regarding human knowledge about the artifact and different capabilities of and information available to the two parties. However, the literature has remained silent on how hidden information – i.e., the principal's lack of knowledge about the agent's characteristics and attitudes before the delegation decision (Bergen et al. 1992; Fayezi et al. 2012) – manifests itself in delegation relationships with nonhuman entities. In purely human delegation relationships, the principal considers the characteristics of the potential agent before weighing whether to delegate a task or perform it themselves (Milewski and Lewis 1997). If the principal knows an agent is better at executing a task, delegation can be beneficial. In this way, complementary team performance can surpass both human and AI performance when performing a task (Fügener et al. 2022). The challenge in AI-to-human delegation is that the AI artifact must estimate the performance of the potential agent before making the delegation decision to evaluate who is better suited to perform the task. This is only possible if there are no information asymmetries between the AI artifact and the person regarding their strengths and weaknesses; instead, the artifact, as the principal, is informed about the human agent's characteristics (E1; 4; 8; 11). Regarding the type of relevant information that serves as parameters for the delegation decision, the research has primarily emphasized AI's ability to estimate the accuracy and probability of the correctness of its output (cf. Fügener et al. 2022; Leibig et al. 2022; E5; 7; 8; 10). This one-dimensional view is insufficient in many use cases, because the AI artifact's accuracy and performance must not be considered in isolation but in conjunction with the human agent (E6; 8; 9; 11). This implies that the two parties' strengths and weaknesses must be considered on an application-specific basis so as to exploit complementarities. It also means that other situation-specific factors should be available to the AI artifact during the delegation decision, such as the cognitive load for the human and their availability (Dubey et al. 2020; E5; 8). However, the essential prerequisite for this is that information flows exist between humans and AI artifacts and that the artifacts can process this information about humans as data input. The difficulty here is defining delegation rules in the AI artifact according to which it performs or delegates tasks. To do this, both parties' strengths and weaknesses must first be defined and quantified,

so that, based on this, the delegation rules' limits can be described (E7; 8; 10; 11).

Going beyond the SLR results, E1, 4, 5, 7, and 8 indicated that – compared to purely human delegation relationships – in AI-to-human delegation, there is not only the problem of information available about the human to the AI artifact but also of the AI's knowledge about itself. Fügener et al. (2022) introduced the term *metaknowledge*, referring to humans' ability to assess their own capabilities. An AI's metaknowledge is critical if AI-to-human delegation is to identify and exploit complementarities in the overall system. It must determine when it is well suited for a task and when it is worse, and when it should delegate tasks to humans to obtain the best possible overall result (E1; 4; 8).

Conflicts of interest

Conflicts of interest are another relevant dimension within PARs, because they can lead to hidden action by the agent and therefore unsuccessful delegation (Arrow 1986). The predominant AI-human conflicts that became apparent in our data were moral and social. An AI artifact as a principal takes decisions based on mathematical operations rather than human cognitive patterns (Carabantes 2020; London 2018; Nabi 2018). On the one hand, this results in a *"cognitive mismatch between the complex mathematical operations performed by (machine learning) algorithms and the type of reasoning used by human beings"* (Carabantes 2020, p. 311). On the other hand, through this abstraction, AI artifacts still have a comparative disadvantage in understanding the complex social and political dynamics that underlie ambiguous decision situations (Jarrahi 2018), which may cause its proposed strategy to violate social constructs and rules in organizations (E2; 9; 12). Further, humans have a sense of morality and can make ethical judgments based on their values and beliefs. While there are attempts to incorporate ethics into AI systems, AI is still unable to make moral judgments in the same way humans can (Yu et al. 2021). This can lead to conflicts of interest when an AI artifact as principal proposes a particular way to achieve a goal that does not conform to a person's personal and/or organizational ethical and moral standards (E3; 6; 9; 10; 11; 12). Thus, E6, 9, and 12 emphasize the need for an ethical and moral framework for an AI artifact as principal that limits the artifact's degrees of freedom in achieving its goals. This lack of a moral framework and social background also justifies the need for human monitoring and possible intervention (Yu et al. 2021).

Further, we identified AI-human conflicts regarding changes in work. While AI artifacts in delegation ownership are often used to optimize work processes' efficiency and augment human capabilities, this also requires a modification of humans' work routines – a situation that people tend to resist (Jarrahi 2018; Kadir and Broberg 2020). In this context E11 stated that *"people tend to want to maintain the status quo, especially in work routines to which they have become accustomed, and then try to behave in such a way that their work routine doesn't change."* This is also reinforced by the fact that new dynamic adjustments can often occur when an AI artifact has delegation ownership (E3; 10; 11; 12). E12 explains this phenomenon by saying that *"many managers often do not critically question their own approaches to task delegation owing to established best practices as well as time and resource constraints"*. AI artifacts, on the other hand, consistently try to achieve the best possible results by considering current circumstances and factors in every decision. However, this conflict of interest may be reduced by building constraints into an artifact's optimization function, such as making task delegation as continuous as possible regarding both time and content (E8; 12).

The third conflict of interest that can arise in AI-to-human delegation is based on economic background. The fear of being replaced makes people act against an AI artifact's interests and instructions (E5; 6; 7; 9; 10; 13); they perceive the ever-advancing development and improvement of AI technologies as a threat because they exceed their own capabilities, making their jobs redundant. This fear then develops into a conflict of interest if a human no longer follows the overriding, shared goal in the cooperation with the AI artifact. For instance, it was stated that this could also lead to false data feedback by humans to degrade an AI artifact's performance (E6; 7; 9; 10; 11). E12 emphasizes that this conflict is also exacerbated by the fact that humans tend to be less afraid of the consequences of not performing a task delegated by an AI as a supervisor than with a human. To prevent this hidden action caused by conflicts of interest, constraints would need to be imposed on human activity, for instance by setting a minimum daily amount of work (E8). Further, outcomes-based incentive contracts that try to align the agent's interests with those of the principal through incentives are an option (Bergen et al. 1992), for instance through further financial incentives or attractive follow-up tasks resulting as a consequence (E11; 12).

Environment and exogenous factors

Our qualitative data showed that, like purely human PARs, AI-to-human delegation is also influenced by external factors. The third aggregate dimension of our results is the external influences and the organizational embedding of the delegation decision that affect a PAR. It was evident from the literature and our data that AI-to-human delegation can only be designed considering the specific use case and social constructions in an organization (Parry et al. 2016; Shrestha et al. 2019). Different requirements must be included in every situation – no one-size-fits-all approach exists (E2; 7; 9; 10; 12). However, we were able to work out high-level dimensions manifested through repeated patterns in our empirical data that need to be considered in delegation design. First, legal regulations on autonomy and explainability limit an artifact's independence and therefore define the roles in AI-to-human delegation to an extent. As noted, it must be ensured that the AI artifacts develop rules only within the de facto legal framework. Further, several laws and regulations require a *right to explanation* to protect the accountability and transparency of automated decisions (e.g., the EU General Data Protection Regulation, GDPR) that require the disclosure of automated decision-making and its underlying logic. This is accompanied by the need to define responsibilities under liability law (Brkan 2019; E3; 6; 13). Currently, AI-to-human delegation must still be preceded by a human, who must not be disregarded, because they determine the overriding target value of the delegation relationship (E4; 6; 7; 8; 9; 10; 13). Further, the design of the AI-to-human delegation will also depend on the potential consequences and sphere of influence of the delegation relationship and their risks. E6 and 7 emphasize the need to design a human-AI system according to the threats posed to which target audience in the event of AI failures and the monetary consequences. The AI's autonomy, its scope for action, and the mechanisms required in each case must be geared to this accordingly. For instance, an artifact with process responsibility for automatic text recognition requires different safety mechanisms to be built in (rather than the single one used in cancer detection).

When describing systems and disclosing a principal's intentions, it became clear that one should also include people's expectations as well as a system's organizational embeddedness (E2; 4; 9; 10; 11; 12), on the one hand, so that the described PAPs are prevented and, on the other hand, the potential of delegation and its benefits can be exploited in the first place. E12 stated in this context that *"the human agent does not have the same expectation towards the human principal as with an AI principal, and this also can cause problems."* There are application areas in which automated decisions are tolerated better than in others where humans still expect a human to be in the leadership role or process ownership position (E6; 13). Although AI-to-human delegation could result in significant advantages here, several factors must be considered in these cases, and the AI-to-human delegation must be aligned accordingly. This could also mean that humans are not informed that an AI artifact instead of a person is now giving them instructions (E2; 3; 6). Thus, besides legal requirements regarding explainability, social expectations must also be considered in the communication.

Human attitude toward the AI artifact as principal

In contrast to purely human delegation relationships, the fact that the principal is no longer human in AI-to-human delegation can also lead to hidden action. Our results have shown that a human's perception of and attitude toward an AI artifact significantly influences whether or not they follow its instructions (Araujo et al. 2020; Bouwer 2022; Cila 2022).

AI artifacts' increasing autonomy induces changes in roles and power for humans, whose attitude toward an artifact is influenced mainly by what changes for them owing to the introduction of the setting in which the artifact is the process owner (E 1; 3; 4; 5; 8; 13). While in the literature, the feeling of giving up control to an algorithm and following its instructions is given as a reason for hidden action (Burton et al. 2020; Colarelli and Thompson 2008), our research shows that more differentiation is needed here. First, it became clear that the occurrence of hidden action by human agents strongly depends on *"who is the beneficiary of the relationship and what is the immediate task that the AI is implementing?"* (E8). In some application areas, humans benefit from handing over process responsibility to artifacts, for instance, in the medical technology sector. There are situations in which the artifact as principal gives instructions to people and delegates tasks to them, that the task-receiving human gladly accepts and even appreciates this because they benefit from them (E1; 4; 10; 11). In other areas, it also happens that this shift automates repetitive tasks and allows people to focus on other, more cognitively demanding or creative tasks (E5; 7; 10). Thus, if the loss of control is accompanied by a human agent being the beneficiary of the relationship, hidden

action does not occur in the first place. However, a further distinction is required here, because it has been found that people are only sometimes aware of their advantages as agents of the system. Further, there are situations in which the primary target variable in AI-to-human delegation is not beneficial to humans; here, other factors define the human agent's actions. In these cases, humans as agents may exhibit algorithm aversion, may not trust a system, or may feel superior to or unappreciated by it. Thus, they may not follow an artifact's instructions or may act against them, even though they are aware of its superior performance (Dietvorst et al. 2015; E1; 5; 6; 7; 9; 11). This can negatively impact on the human-AI PAR, since trust influences human reliance on automation and is indispensable for their acceptance of AI-made decisions (Lee and See 2004).

To address these problems, the literature has mainly referred to explainability (cf. Berente et al. 2021; Burton et al. 2020; Wilson and Daughtery 2018). However, as mentioned, to date, providing humans with accurate and understandable explanations about the functioning of intelligent systems is only possible to a limited extent. The debate about explainability is vast and essential, especially against a legal and governance background, but is secondary when looking at the direct human-AI relationship. Informing a human about an artifact's way of working and its underlying delegation rules will not reduce hidden action in every situation. It cannot be assumed that all people want to understand how a system works or how it came to a decision; often, the explanations are perceived as too complex or uninteresting (E2; 3; 6; 10; 12). E4 and 8 emphasize that giving human agents precise explanations about an AI artifact's behavior can even negatively affect their attitude toward it and can lead to deterrence and hidden action. Instead, to prevent hidden action and improve people's attitude toward AI artifacts, it is more important to show a person an artifact's performance and advantages instead of trying to explain its functioning (E1; 3; 5; 6; 11; 13). In many situations, people accept an artifact's instruction not because they know how it took that decision, but because they have experienced it and have seen it perform well (E1; 3; 4; 5; 6; 11). Further, E1, 2, 6, 8, 9, and 13 emphasize the need to adapt AI communication to the situation at hand, but above all, to make it easy to understand and straightforward, leaving as little room as possible for human interpretation. The goal should also be to make communication as human as possible, so that AI artifacts can empathize with humans and can respond interactively to them (E2, 3). Moreover, communication within AI-to-human delegation should be human- and situation-specific when preventing or reducing hidden action. Explanations, transparency, and how the instruction is imparted need to be adapted to every person's characteristics (Vössing et al. 2022). According to E9, two important questions to ask when communicating the task from the AI artifact to the human are, *"Who is the user? What is the situation?"*. For instance, artifacts should consider the human agent's expertise, job position, and attitude toward the AI, for instance, algorithm aversion, fear of being replaced, lack of trust, the feeling of not being appreciated. This is especially significant, because how autonomous systems are explained and interact with humans dramatically impacts on their behavior (cf. Berger et al. 2021; Burton et al. 2020; E1; 8; 9; 12; 13). It must be evaluated when, for instance, communication needs to be more fact-based or more empathetic, as well as what information is supplied when a delegation task is transferred. It has been shown that the more people are experts in their field (e.g., ML software developers), the more they feel superior to the AI artifact and, therefore, do not follow its instructions but their intuition and gut feeling (Mahmud et al. 2022; E9; 13). The same applies to people's technical understanding. While nontechnical people often have an automation bias (i.e., they rely entirely on a technology), technically-versed users more often question an artifact's decisions and expect logical explanations and evidence (Goddard et al. 2014; E6; 8; 12).

In line with the literature, it was also stated that people sometimes need to be made to feel that they are still in control or required at specific points in the process, even if this is not theoretically the case (Burton et al. 2020; Colarelli and Thompson 2008; E5), because humans lose sovereignty over the process to a technological entity to which they must subordinate. However, to trust and rely on an artifact's judgment, they need to feel that they have control over the algorithm (Burton et al. 2020; Colarelli and Thompson 2008; Fügener et al. 2022; E4; 5; 6). Further, E4 and 5 state that the consideration of algorithm aversion, trust, etc. within AI-to-human delegation should even be extended to include other situation-specific factors, because people sometimes act against a system for different reasons, for instance, if they are afraid of possible consequences or have a bad conscience.

Discussion and Conclusion

Our work has pushed the scientific frontier beyond primarily human-driven task delegation by addressing the increasing autonomy of AI artifacts and the transition from delegation ownership to it. We contribute to the understanding of PARs in AI-to-human delegation by examining and extending the concepts of PAT into the context of this new delegation setting. This explorative study shows that our assumption that PAT is also applicable to AI-to-human delegation holds true. The two PAPs in AI-to-human delegation - hidden information and hidden action - align with the general concept proposed in the literature. As in purely human delegation relationships, PAPs in AI-to-human delegation arise owing to information asymmetries and conflicts of interest and are influenced by exogenous factors and the environment (Bergen et al. 1992; Keil 2005; Peterson 1993). Expanding the body of knowledge, we have also uncovered new causes of PAPs that are specifically caused by AI artifacts as principals, namely the attitude of the human agent toward the AI artifact. For example, the human perception of the AI artifact as a capable, superior team partner and the confidence in its capabilities and performance largely determine whether or not hidden action occurs.

We demonstrate the need for new solution mechanisms for arising PAPs in AI-to-human delegation. Research in human-AI collaboration has already highlighted human centricity as an important element in designing AI applications (Candrian and Scherer 2022; Hemmer et al. 2021; Yu et al. 2021). While our study allows making statements on the effect of AI artifacts as a principal and the aspects to be considered, it has particularly become clear that human centricity is of overriding relevance in this delegation relationship. Although the AI artifact now holds the delegation ownership, the human remains the entity to which the delegation relationship must be oriented. Thus, we argue for a distinction with respect to the necessity of explainable AI decisions. While previous research considers explainable AI essential for successful human-AI collaboration, our data has shown that the use of explanations in AI-to-human delegation should also be adapted to the respective human to avoid hidden action and make delegation successful.

Furthermore, our results have shown that AI-to-human delegation changes the principal's responsibilities. Until now, the human has been the responsible entity for receiving, evaluating, and possibly delegating the task to an intelligent agent, as well as for ensuring the correctness of the final result, thus assuming technical and legal responsibility for the overall process. By letting the AI artifact decide whether to perform the task itself or delegate it, there may no longer be any human involvement in the process, making the artifact accountable for the process outcome. Against the background of the worldwide common legal view that only human entities can be responsible (Benbya et al. 2020; Loi and Spielkamp 2021), the change from "human-to-AI" to "AI-to-human" creates a separation of the levels of responsibility: The AI artifact as principal becomes accountable, but the legal responsibility must be externalized. For a fully autonomous delegation by the AI artifact, there must be a human who takes over responsibility, for example, the human agent or another entity. However, the more responsibility is ensured by measures in the process (e.g., human quality control of the final decision at the end of the delegation process), the lower the efficiency gain of AI-to-human delegation compared to human-to-AI is. Future research should investigate how to achieve this separation of levels of responsibility in AI-to-human delegation without compromising the efficiency of the delegation process, e.g., through a design science research approach.

From a practical perspective, the study goal was to reduce the uncertainty regarding the setting of autonomous AI artifacts with delegation ownership giving instructions to humans. Thus, we examined challenges and requirements, presenting a selection of effects that are relevant for AI-led delegation processes in practice. We have extended the delegation patterns and rules of the AI artifacts applied to date in practice by showing which further factors must be considered by the AI in the delegation decision to make the collaboration as profitable as possible for all parties involved. These results also provide guidance to developers regarding required and suitable information flows between the two delegation entities. We have also contributed to the understanding of how synergies between humans and AI artifacts can be created in the delegation relationship. The different skills and approaches to problems between AI artifacts as principals and human agents can be used not only as a challenge but also to complement each other (Hemmer et al. 2021), for instance by having humans act as data providers.

Our research has limitations that may stimulate future research. Notably, our interview study sample size could limit the generalizability of our findings. However, we are confident that we captured the most relevant constructs, as we had reached a certain degree of saturation toward the end of the data analysis. Also, we approached new constructs exploratively and expanded scientific knowledge. Further, we used

qualitative empirical research methods, and obtained the dataset used to generate the results through semi-structured interviews with experts. Although we used numerous techniques to improve internal and external validity as well as objectivity (Lombard et al. 2002), the coding of the data was based on subjective judgment. Finally, we encourage researchers to further investigate AI-to-human delegation as the significant impact of rapidly advancing AI technology on humans suggests a need to explore this form of collaboration. For instance, the topic should be explored in depth through case studies to provide unique and novel insights into the tensions between AI and humans in real-world settings. Future research could also focus on developing solution mechanisms for our identified problems and challenges in AI-to-human delegation where the classical solutions within PAT are no longer sufficient. Further, our consideration of only 1:1 relationships could be extended to 1:n or n:n relationships between AI artifacts and humans.

References

Abbasi, M. F., Bilal, M., and Rasheed, K. 2022. "Role of Human Intuition in AI Aided Managerial Decision Making: A Review," in *2022 International Conference on Decision Aid Sciences and Applications (DASA),* Chiangrai, Thailand, IEEE, pp. 713-718.

Ågerfalk, P. J. 2020. "Artificial intelligence as digital agency," *European Journal of Information Systems* (29:1), pp. 1-8.

Araujo, T., Helberger, N., Kruikemeier, S., and Vreese, C. H. de. 2020. "In AI we trust? Perceptions about automated decision-making by artificial intelligence," *AI & Society* (35:3), pp. 611-623.

Arrow, K. J. 1986. "Agency and the market," in *Handbook of Mathematical Economics*, K. J. Arrow and M. D. Intriligator (eds.), Elsevier, pp. 1183-1195.

Baird, A., and Maruping, L. M. 2021. "The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts," *MIS Quarterly* (45:1), pp. 315-341.

Bawack, R., Wamba, S. F., and Carillo, K. D. 2019. "From IT to AI Artifact: Implications for IS Research on AI Adoption and Use," in *Proceedings of the Twenty-fourth DIGIT Workshop,* Munich, Germany.

Beer, J. M., Fisk, A. D., and Rogers, W. A. 2014. "Toward a framework for levels of robot autonomy in human-robot interaction," *Journal of human-robot interaction* (3:2), pp. 74-99.

Benbya, H., Davenport, T. H., and Pachidi, S. 2020. "Artificial Intelligence in Organizations: Current State and Future Opportunities," *MIS Quarterly Executive* (19:4), pp. ix-xxi.

Berente, N., Gu, B., Recker, J., and Santhanam, R. 2021. "Managing Artificial Intelligence," *MIS Quarterly* (45:3), pp. 1433-1450.

Bergen, M., Dutta, S., and Walker, O. C., JR. 1992. "Agency Relationships in Marketing: A Review of the Implications and Applications of Agency and Related Theories," *Journal of Marketing* (56:3), pp. 1-24.

Berger, B., Adam, M., Rühr, A., and Benlian, A. 2021. "Watch Me Improve - Algorithm Aversion and Demonstrating the Ability to Learn," *Business & Information Systems Engineering* (63:1), pp. 55-68.

Berle, A. A., and Means, G. C. 1932. *The modern corporation and private property*, New York: The Macmillan Company.

Bogner, A., and Menz, W. 2009. "The Theory-generating Expert Interview: Epistemological Interest, Forms of Knowledge, Interaction," in *Interviewing Experts*, A. Bogner, B. Littig and W. Menz (eds.), London: Palgrave Macmillan UK, pp. 43-80.

Borch, C. 2022. "Machine learning, knowledge risk, and principal-agent problems in automated trading," *Technology in Society* (68:3), p. 101852.

Bouwen, G. A. 2006. "Grounded Theory and Sensitizing Concepts," *International Journal of Qualitative Methods* (5:3), pp. 12-23.

Bouwer, A. 2022. "Under Which Conditions Are Humans Motivated to Delegate Tasks to AI? A Taxonomy on the Human Emotional State Driving the Motivation for AI Delegation," in *Marketing and Smart Technologies*, J. L. Reis, E. P. López, L. Moutinho and J. P. M. dos Santos (eds.), Singapore: Springer, pp. 37-53.

Brkan, M. 2019. "Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond," *International Journal of Law and Information Technology* (27:2), pp. 91-121.

Burton, J. W., Stein, M.-K., and Jensen, Blegind, Tina. 2020. "A systematic review of algorithm aversion in augmented decision making," *Journal of Behavioral Decision Making* (33:2), pp. 220-239.

Candrian, C., and Scherer, A. 2022. "Rise of the machines: Delegating decisions to autonomous AI," *Computers in Human Behavior* (134:4), p. 107308.

Carabantes, M. 2020. "Black-box artificial intelligence: an epistemological and critical analysis," *AI & Society* (35:2), pp. 309-317.
Charmaz, K. 2000. "Grounded Theory: Objectivist and Constructivist Methods," in *The Handbook of Qualitative Research*, N. K. Denzin and Y. S. Lincoln (eds.), Thousand Oaks, CA: Sage Publications, pp. 249-291.
Chen, J. Y. C., and Barnes, M. J. 2014. "Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues," *IEEE Transactions on Human-Machine Systems* (44:1), pp. 13-29.
Cila, N. 2022. "Designing Human-Agent Collaborations: Commitment, responsiveness, and support," in *CHI Conference on Human Factors in Computing Systems*, S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson and K. Yatani (eds.), New York, USA: ACM, pp. 1-18.
Colarelli, S. M., and Thompson, M. 2008. "Stubborn Reliance on Human Nature in Employee Selection: Statistical Decision Aids Are Evolutionarily Novel," *Industrial and Organizational Psychology* (1:3), pp. 347-351.
Corbin, J., and Strauss, A. 2015. *Basics of Qualitative Research: Techniques and procedures for developing grounded theory*, Los Angeles, California: Sage Publications.
Corbin, J. M., and Strauss, A. 1990. "Grounded theory research: Procedures, canons, and evaluative criteria," *Qualitative Sociology* (13), pp. 3-21.
Dattathrani, S., and De', R. 2023. "The Concept of Agency in the Era of Artificial Intelligence: Dimensions and Degrees," *Information Systems Frontiers* (25:1), pp. 29-54.
Dellermann, D., Calma, A., Lipusch, N., Weber, T., and Weigel, S. & Ebel, P. 2019. "The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems.," in *Hawaii International Conference on System Sciences (HICSS)*, Hawaii, USA.
Dietvorst, B. J., Simmons, J. P., and Massey, C. 2015. "Algorithm aversion: people erroneously avoid algorithms after seeing them err," *Journal of experimental psychology. General* (144:1), pp. 114-126.
Dubey, A., Abhinav, K., Jain, S., Arora, V., and Puttaveerana, A. 2020. "HACO: A Framework for Developing Human-AI Teaming," in *Proceedings of the 13th Innovations in Software Engineering Conference*, S. Jain, A. Gupta, D. Lo, D. Saha and R. Sharma (eds.), New York, USA: ACM, pp. 1-9.
Eisenhardt, K. M. 1989. "Agency Theory: An Assessment and Review," *Academy of Management Review* (14:1), pp. 57-74.
Eisenhardt, K. M., and Graebner, M. E. 2007. "Theory building from cases: Opportunities and challenges," *Academy of Management Journal* (50:1), pp. 25-32.
Emirbayer, M., and Mische, A. 1998. "What is Agency?" *The American Journal of Sociology* (103:4), pp. 962-1023.
Farooq, U., and Grudin, J. 2016. "Human-Computer Integration," *Interactions* (23:6), pp. 26-32.
Fayezi, S., O'Loughlin, A., and Zutshi, A. 2012. "Agency theory and supply chain management: a structured literature review," *Supply Chain Management* (17:5), pp. 556-570.
Fuchs, A., Passarella, A., and Conti, M. 2022. "A Cognitive Framework for Delegation Between Error-Prone AI and Human Agents," *ArXiv, abs/2204.02889v3*.
Fügener, A., Grahl, J., Gupta, A., and Ketter, W. 2022. "Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation," *Information Systems Research* (33:2), pp. 678-696.
Gioia, D. A., Corley, K. G., and Hamilton, A. L. 2013. "Seeking Qualitative Rigor in Inductive Research," *Organizational Research Methods* (16:1), pp. 15-31.
Glaser, B. G. 1978. *Theoretical Sensitivity: Advances in the Methodology of Grounded Theory*, Mill Valley, CA: The Sociology Press.
Goddard, K., Roudsari, A., and Wyatt, J. C. 2014. "Automation bias: empirical results assessing influencing factors," *International journal of medical informatics* (83:5), pp. 368-375.
Goldbach, C., Kayar, D., Pitz, T., and Sickmann, J. 2019. "Transferring decisions to an algorithm: A simple route choice experiment," *Transportation Research Part F: Traffic Psychology and Behaviour* (65:6), pp. 402-417.
Goodman, L. A. 1961. "Snowball Sampling," *The Annals of Mathematical Statistics* (32:1), pp. 148-170.
Griffiths, N. 2005. "Task delegation using Experience-Based Multi-Dimensional Trust," in *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems (AAMAS '05),* New York, NY, USA, pp. 489-496.
Guggenberger, T., Lämmermann, L., Urbach, N., Walter, A., and Hofmann, P. 2023. "Supplementary material for the publication 'Task delegation from AI to humans: A principal-agent perspective'". *Zenodo, https://zenodo.org/record/8319026*.

Gur, N., and Bjørnskov, C. 2017. "Trust and delegation: Theory and evidence," *Journal of Comparative Economics* (45:3), pp. 644-657.
Harms, P. D., and Han, G. 2019. "Algorithmic Leadership: The Future is Now," *Journal of Leadership Studies* (12:4), pp. 74-75.
Heaslip, G., and Kovács, G. 2018. "Examination of service triads in humanitarian logistics," *The International Journal of Logistics Management* (30:7), pp. 595-619.
Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., and Satzger, G. 2022. "On the Effect of Information Asymmetry in Human-AI Teams," *ArXiv, abs/2205.01467v1*.
Hemmer, P., Schemmer, M., Vössing, M., and Kühl, N. 2021. "Human-AI Complementarity in Hybrid Intelligence Systems," in *Twenty-fifth Pacific Asia Conference on Information Systems*, Dubai, UAE.
Jarrahi, M. H. 2018. "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," *Business Horizons* (61:4), pp. 577-586.
Jennings, N. R., Moreau, L., Nicholson, D., Ramchurn, S., Roberts, S., Rodden, T., and Rogers, A. 2014. "Human-agent collectives," *Communications of the ACM* (57:12), pp. 80-88.
Jensen, M. C., and Meckling, W. H. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *Journal of Financial Economics* (4:3), pp. 305-360.
Kadir, B. A., and Broberg, O. 2020. "Human well-being and system performance in the transition to industry 4.0," *International Journal of Industrial Ergonomics* (76), p. 102936.
Keil, P. 2005. "Principal Agent Theory and its application to analyze outsourcing of software development," *ACM SIGSOFT Software Engineering Notes* (30), 1,5.
Kim, E.-S. 2020. "Deep learning and principal–agent problems of algorithmic governance: The new materialism perspective," *Technology in Society* (63:4), p. 101378.
Lazányi, K. 2018. "Are we ready for self-driving cars - a case of Principal-Agent Theory," in *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI),* Timisoara, Romania, IEEE, pp. 251-254.
Lee, I., and Shin, Y. J. 2018. "Fintech: Ecosystem, business models, investment decisions, and challenges," *Business Horizons* (61:1), pp. 35-46.
Lee, J. D., and See, K. A. 2004. "Trust in Automation: Designing for Appropriate Reliance," *Human factors* (46:1), pp. 50-80.
Leibig, C., Brehmer, M., Bunk, S., Byng, D., Pinker, K., and Umutlu, L. 2022. "Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis," *The Lancet Digital Health* (4:7), e507-e519.
Loi, M., and Spielkamp, M. 2021. "Towards Accountability in the Use of Artificial Intelligence for Public Administrations," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, M. Fourcade, B. Kuipers, S. Lazar and D. Mulligan (eds.), New York, USA: ACM, pp. 757-766.
Lombard, M., Snyder-Duch, J., and Bracken, C. C. 2002. "Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability," *Human Communication Research* (28:4), pp. 587-604.
London, A. J. 2018. "Groundhog Day for Medical Artificial Intelligence," *The Hastings Center report* (48:3).
Mahmud, H., Islam, A. N., Ahmed, S. I., and Smolander, K. 2022. "What influences algorithmic decision-making? A systematic literature review on algorithm aversion," *Technological Forecasting and Social Change* (175:49), p. 121390.
Markus, M. L. 2017. "Datification, Organizational Strategy, and IS Research: What's the Score?" *The Journal of Strategic Information Systems* (26:3), pp. 233-241.
McClure, J. 2014. "Conceptual Parallels between Philosophy of Science and Cognitive science: Artificial Intelligence, Human Intuition, and Rationality," *Aporia* (24:1), pp. 39-49.
Milewski, A. E., and Lewis, S. H. 1997. "Delegating to Software Agents," *International Journal of Human-Computer Studies* (46:4), pp. 485-500.
Morris, A. T., Maddalon, J. M., and Miner, P. S. 2020. "On the Moral Hazard of Autonomy," in *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC),* San Antonio, USA, IEEE, pp. 1-9.
Myers, M. D., and Newman, M. 2007. "The qualitative interview in IS research: Examining the craft," *Information and Organization* (17:1), pp. 2-26.
Nabi, J. 2018. "How Bioethics Can Shape Artificial Intelligence and Machine Learning," *Hastings Center Report* (48:5), pp. 10-13.
Noy, C. 2008. "Sampling Knowledge: The Hermeneutics of Snowball Sampling in Qualitative Research," *International Journal of Social Research Methodology* (11:4), pp. 327-344.

Padgett, D. K. 2003. "Coming of age: Theoretical thinking, social responsibility, and a global perspective in qualitative research," in *The Qualitative Research Experience*, D. K. Padgett (ed.), Belmont, USA: Wadsworth/Thomson Learning, pp. 297-315.
Parry, K., Cohen, M., and Bhattacharya, S. 2016. "Rise of the Machines," *Group & Organization Management* (41:5), pp. 571-594.
Pasquale, F. 2015. *The black box society*: *The secret algorithms that control money and information*, Cambridge, Mass., London: Harvard University Press.
Pavlou, P. A., Liang, H., and Xue, Y. 2007. "Understanding and Mitigating Uncertainty in Online Exchange Relationships: A Principal-Agent Perspective," *MIS Quarterly* (31:1), pp. 105-136.
Peeters, M. M. M., van Diggelen, J., van den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., and Raaijmakers, S. 2021. "Hybrid collective intelligence in a human–AI society," *AI & Society* (36:1), pp. 217-238.
Peterson, T. 1993. "Recent Developments in: The Economics of Organization: The Principal-Agent Relationship," *Acta Sociologica* (36:3), pp. 277-293.
Poth, S., and Selck, T. J. 2009. "Principal Agent Theory and Artificial Information Asymmetry," *Politics* (29:2), pp. 137-144.
Rai, A., Constantinides, P., and Sarker, S. 2019. "Next-Generation Digital Platforms: Toward Human-AI Hybrids," *MIS Quarterly* (43:1), pp. iii-x.
Schneider, D. 1987. "Agency Costs and Transaction Costs: Flops in the Principal-Agent-Theory of Financial Markets," in *Agency Theory, Information, and Incentives*, G. Bamberg and K. Spremann (eds.), Heidelberg, Germany: Springer.
Schuetz, S., and Venkatesh, V. 2020. "The rise of human machines: How cognitive computing systems challenge assumptions of user-system interaction," *Journal of the Association for Information Systems* (21:2), pp. 460-482.
Shapiro, S. P. 2005. "Agency Theory," *Annual Review of Sociology* (31:1), pp. 263-284.
Shrestha, Y. R., Ben-Menahem, S. M., and Krogh, G. von. 2019. "Organizational Decision-Making Structures in the Age of Artificial Intelligence," *California Management Review* (61:4), pp. 66-83.
Steffel, M., Williams, E. F., and Perrmann-Graham, J. 2016. "Passing the buck: Delegating choices to others to avoid responsibility and blame," *Organizational Behavior and Human Decision Processes* (135), pp. 32-44.
Tate, W. L., Ellram, L. M., Bals, L., Hartmann, E., and van der Valk, W. 2010. "An Agency Theory perspective on the purchase of marketing services," *Industrial Marketing Management* (39:5), pp. 806-819.
Vincent, V. U. 2021. "Integrating intuition and artificial intelligence in organizational decision-making," *Business Horizons* (64:4), pp. 425-438.
Vössing, M., Kühl, N., Lind, M., and Satzger, G. 2022. "Designing Transparency for Effective Human-AI Collaboration," *Information Systems Frontiers* (25:2), p. 954.
Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly* (26:2), pp. xiii-xxviii.
Wesche, J. S., and Sonderegger, A. 2019. "When computers take the lead: The automation of leadership," *Computers in Human Behavior* (101:12), pp. 197-209.
Wilson, H. J., and Daughtery, P. R. 2018. "Collaborative Intelligence: Humans and AI are joining forces," *Harvard Business Review* (96:4), pp. 114-123.
Xu, W., Dainoff, M. J., Ge, L., and Gao, Z. 2023. "Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI," *International Journal of Human–Computer Interaction* (39:3), pp. 494-518.
Yu, B., Vahidov, R., and Kersten, G. E. 2021. "Acceptance of technological agency: Beyond the perception of utilitarian value," *Information & Management* (58:7), p. 103503.
Yukl, G., and Fu, P. P. 1999. "Determinants of delegation and consultation by managers," *Journal of Organizational Behavior* (20:2), pp. 219-232.