See No Evil, Hear No Evil: How Users Blindly Overrely on Robots with Automation Bias

Ruth Stock-Homburg
Technical University Darmstadt, ruth.stock-homburg@tu-darmstadt.de

Mai Anh Nguyen
Technical University of Darmstadt, mangn.tud@gmail.com

See No Evil, Hear No Evil: How Users Overrely on Robots with Automation Bias

Completed Research Paper

Ruth Stock-Homburg
Technical University of Darmstadt
Hochschulstr. 1, 64289 Darmstadt, Germany
rsh@bwl.tu-darmstadt.de

Mai Anh Nguyen
Technical University of Darmstadt
Hochschulstr. 1, 64289 Darmstadt, Germany
mangn.tud@gmail.com

Abstract

Recent developments in generative artificial intelligence show how quickly users carelessly adhere to intelligent systems, ignoring systems' vulnerabilities and focusing on their superior capabilities. This is detrimental when system failures are ignored. This paper investigates this mindless overreliance on systems, defined as automation bias (AB), in human-robot interaction. We conducted two experimental studies (N1 = 210, N2 = 438) with social robots in a corporate setting to investigate psychological mechanisms and influencing factors of AB. Particularly, users experience perceptual and behavioral AB with the robot that is enhanced by robot competence depending on task complexity and is even stronger for emotional than analytical tasks. Surprisingly, robot reliability negatively affected AB. We also found a negative indirect-only mediation of AB on robot satisfaction. Finally, we provide implications for the appropriate use of robots to prevent employees from using them as a self-sufficient system instead of a supporting system.

Keywords: automation bias, overreliance, human-robot interaction, authority

Introduction

Most of us are accustomed to using a navigation system to guide us to a specific location. Over time, we stop questioning the navigation system's suggestions and rather mindlessly trust its performance. This mindlessness can be frustrating when the system erroneously leads us to a wrong route or detour. Mindlessly accepting incorrect information can lead to bad decisions and poor outcomes. The example illustrates the cognitive bias *automation bias (AB)* that is the "tendency to over-rely on automation" (Goddard et al., 2012, p. 121). Positive experiences from the overreliance on automation thus reinforce a mindless behavior that reduces critical reflection of automation information and increases the likelihood of errors (Parasuraman & Manzey, 2010). AB can also occur during collaborative human-robot interactions (HRIs), in which employees, to some extent, mindlessly rely on robots. Ongoing advances in machine learning and AI-based programs (e.g., ChatGPT) accelerate robotization in the corporate environment so that robots already make up to 51% of the US economy (Smids et al., 2020). These robots have equal or even superior abilities to humans, which enable them to assist humans with increasingly complex and non-routinized cognitive tasks (Fleming, 2019). In particular, they perform by interacting in a natural and easy-to-understand way or providing services to employees, such as coordination, information search, or decision-making support (Kirby et al., 2010; Smids et al., 2020). As a result, *organizational robots* that are artificially intelligent and embodied machines that can execute tasks that typically require human intelligence, will have a great impact on the working world (Smids et al., 2020). With these capabilities, unlike self-service technologies, they can mimic social characteristics, which makes them *social organizational robots* with automated social presence (Belanche et al., 2020). This is vital because service robots are increasingly present in the workplace as real anthropomorphized (i.e., humanized) interaction

partners and less as functional automations (Yam et al., 2020). Thus, investigating AB in the context of HRI enables us to embed AB in a social context and includes potential effects of anthropomorphism and automated social presence. Although these robots can drive benefits (e.g., cost savings, productivity, and innovation), they can also be disadvantageous if being misused (Parasuraman & Riley, 1997). This occurs when employees do not use the robot in the predefined way as an assistive system, but as a replacement system for own thinking (Parasuraman & Manzey, 2010). Especially, in a work environment with increasing workload and time pressure, bounded rationality is reinforced, which can affect decision-making quality (DQ) and form the basis of AB (Simon et al., 2000). The extent to which AB leads to DQ can be observed in various organizational settings as demonstrated by Amazon's biased AI recruiting tool (Parikh, 2021). Thus, incorrect information provided by social organizational robots that is not recognized as such can lead to poor decisions (McKibbon & Fridsma, 2006). Especially in organizations, DQ is crucial and can have severe negative consequences if it is low. Consequently, integrated research on AB and HRI in organizations is needed to understand AB during HRI and its outcomes and to optimize human-robot collaboration (HRC). Thus, we address following research questions with this paper:

RQ1: *To what extent does AB occur in the context of organizational HRI on a perceptual and behavioral level?* Due to the latent nature of AB, we investigate it within the perception of the user to understand if and how it occurs from a psychological perspective (via a perceived reduction of one's own abilities relative to the robot). We further investigate how AB is made explicit in behavior (via compliance with the robot's advice) to gain a deeper understanding for the bias. With this dual approach, we can attribute AB to attitudinal and cognitive distortion towards robots and not other factors such as inattention. Therefore, we compare AB with reliance on the human as the control group.

RQ2*: What moderating factors affect the strength of AB during HRI in organizations?* Understanding the moderating factors of AB helps to understand how to influence the strength of AB. As moderating factors, we examine robot-related factors (competence and reliability) and task-related factors (complexity and type). Knowledge of this can contribute to a more comprehensive understanding of how AB differs across various work settings or robot characteristics (e.g., differences of AB with an error-prone robot in a analytical task such as financial forecasts and emotional task such as recruiting). This in turn, can be used to adapt the social organizational robot to users or tasks so that the overall HRC is optimized, and negative effects of AB are reduced or prevented.

RQ3*: To what extent does AB toward a social organizational robot affect employee satisfaction with it?* Satisfaction with the robot as an outcome of AB is included in this study to understand proximal consequences of AB on the overall HRC. Insights on the impact of AB on user satisfaction may reveal more about the ambiguity of AB, which either shows in a positive evaluation (indicating use) or a negative evaluation (indicating misuse) (Parasuraman & Riley, 1997). Therefore, we investigate mediating effects of AB on robot satisfaction.

To answer these questions, we conducted an experimental study in a real-world corporate setting (study 1) and an online experimental study via Amazon Mechanical Turk USA (study 2). In study 1, we focused on the natural occurrence of AB at a perceptual and behavioral level, providing a basic understanding of AB. Study 2 examined the systematic elicitation of AB through the manipulation of robot reliability and included performance change measurements. Drawing on automation authority and superiority theory and diffusion of responsibility theory, we shed light on mechanisms, expression, influencing factors, and effects of AB.

Literature Review

Research on AB originates from aircraft and relies primarily on behavioral performance as indicators of AB (Goddard et al., 2012). Particularly, during tasks that require high attention (e.g., monitoring or diagnosis), people tend to overrely on the faulty automation, which in turn results in poor performance. Besides, reinforcing contingency factors such as automation reliability (Parasuraman et al., 1993), task load (Lyell & Coiera, 2017), trust in automation (Yeh & Wickens, 2001), and own responsibility (Mosier & Skitka, 1996; Shah & Bliss, 2017) have been identified to influence AB. These findings have been replicated in several settings, including military, healthcare, and aircraft (Goddard et al., 2012). Thus, AB occurs in different situations on a behavioral level and is affected by automation-related, task-related, and user-related factors. Despite these findings, the phenomenon of AB is obscure and lacks understanding (Goddard et al., 2012). Our literature review revealed that current research only measures AB at the behavioral level through poor

performance. This operationalization may be insufficient, as poor performance can occur for various reasons (e.g., inattention or delayed action) (Moray et al., 2000). Thus, most studies determine AB through behavioral outcomes (i.e., commission or omission errors), but do not test the proposed dominant use of automated cues (Moray et al., 2000). One way to address this shortcoming is to include a "normative model of optimal ("eutactic") behavior" (Bahner et al., 2008, p. 689) for comparison with biased behavior. Consequently, a more comprehensive measure is lacking that includes both underlying psychological mechanisms, represented at the perceptual level of AB, and user actions, representing the behavioral level. In addition, most of the AB studies include simulations and do not reflect real-world conditions, which limits their generalizability (Goddard et al., 2012). Moreover, existing studies are mainly based on a biased expert sample with specific working groups (e.g., pilots or physicians) (Goddard et al., 2012; Parasuraman & Manzey, 2010). Thus, AB has been analyzed foremost in safety-critical situations where overreliance on faulty automations can be detrimental. These findings may not apply to daily life with lower-risk situations such as at work, where the situation is more low-threshold, requires less cognitive effort and decision-makers have less expertise (i.e., laypeople) and responsibility (i.e., milder consequences).

We address these shortcomings and contribute to the research on HRI in several ways. First, we extend the scope of AB research to the context of HRI by analyzing AB as a systematic bias during HRI in a corporate setting. Although there are some studies in HRI, which address the topic of reliance, they primarily focus on blind trust in robots (Robinette et al., 2016), which should be distinguished from AB. This distinction is vital as it influences but does not fully define reliance (Lee & See, 2004). Second, we extend AB literature with insights into underlying mechanisms by drawing on psychological theories. Based on this, we develop a measurement for AB and include control groups to clearly distinguish AB from normative behavior. Third, we uncover robot-related and task-related moderators of AB, which helps to identify measures to influence AB in decision-making processes. Fourth, we analyze AB in an experimental study in a real-life setting that provides an experimental study design realism and a sense of immersion (Vanden Abeele & Postma-Nilsenova, 2018). This increases external and internal validity and thus allows generalizability and causal inference (Aguinis & Bradley, 2014). Fifth, we show that employees are exposed to AB in low-risk situations, suggesting that AB concerns a great and diverse sample. This highlights the everyday relevance of AB in work situations and reinforces the generalizability of findings. Finally, we investigate the mediating effect of AB on satisfaction with the robot to provide insights into the adequate use of social organizational robots and to contribute to reinforcing successful HRCs. This addresses the research gap on collaborative technologies and contributes to current opportunities in computer science (Benbya et al., 2021).

Definition and Conceptualization of Automation Bias

In general, AB represents mental shortcuts in human cognition that are used to reduce a person's cognitive effort (Simon et al., 2000). The term AB was originally coined in aviation and is defined as the "tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing" (Mosier & Skitka, 1996, p. 205), which involves the assumption of automation infallibility (McKibbon & Fridsma, 2006). AB includes commission errors and omission errors (Manzey et al., 2006). *Commission errors* deal with the phenomenon of users inappropriately following the system's recommendations "without verifying it against other available information or despite contradictions" (Mosier et al., 2001, p. 2). *Omission errors* are failures to adequately respond to critical situations, when not prompted by the automation (Manzey et al., 2006). Thus, AB can lead to errors based on insufficient information analysis and domination of automation advice. This has been explained by the general belief in the infallibility of the automation, reduced vigilance, reluctance to pursue conflicting evidence against automation information, and diffusion of responsibility that are complementary. Together, they form a kind of compensatory mechanism to reduce the overall workload (Bahner et al., 2008; McKibbon & Fridsma, 2006).

AB is to be distinguished from related terms in computer science, namely algorithm appreciation and complacency. *Algorithm appreciation* describes the preference for algorithmic advice over human advice, when no algorithmic error occurs (Logg et al., 2019). The main difference from AB is that algorithm appreciation decreases or disappears after algorithm error, while AB occurs despite automation error, revealing commission error (Logg et al., 2019). *Complacency* is a "psychological state characterized by a low index of suspicion" (Wiener, 1981, p. 117) that can lead to reduced consideration of technological recommendations. Particularly, complacency focuses on user attention, whereas AB focuses on behavioral and performance outcomes (Parasuraman & Manzey, 2010).

Here, we define AB as a heuristic shortcut in human cognition during HRI, reflected by the reduction of own perceived responsibility, control, power, and autonomy, implying a reduction of own perceived abilities. AB is not rooted in one's inability, but in one's underestimation and simultaneous overestimation of the robot. The derivation of this definition is based on the theoretical basis of AB discussed next.

Theoretical Background: Imbalance of Responsibility and Authority

We rely on diffusion of responsibility theory (DRT) and automation authority and superiority theory (AAST) to explain underlying mechanisms of AB. DRT suggests that people in company feel less personal responsibility, i.e., decision-makers' sense of ownership for an outcome or actions (Botti & McGill, 2006) than when they are alone (Darley & Latané, 1968). The shift in perceived responsibility occurs from self to others (Forsyth et al., 2002). Accordingly, during HRI, responsibility shifts from the user to the robot, with the user mindlessly relying on the robot. Consistently, extant AB literature draws on the DRT, suggesting that AB-prone users feel less responsible when using automation (McKibbon & Fridsma, 2006; Shah & Bliss, 2017; Skitka et al., 2000). As a result, they feel less obligated to make an effort. Furthermore, Jörling et al. (2019) showed that this is reinforced by a high level of automation autonomy because it decreases one's perceived behavioral control and perceived responsibility. Skitka et al. (2000) even found a strong positive correlation between diffusion of responsibility and errors resulting from AB. Finally, the DRT can be used to identify the assignment of responsibility as a fundamental mechanism in the occurrence of AB. However, it does not provide insights into why the shift in responsibility occurs. These insights can be gained from signal detection research, more specifically AAST that explains how automated aid signals are perceived and how users comply with them (Mosier & Skitka, 1996).

AAST postulates that automations are perceived as powerful agents with greater capabilities than humans, which can unfold, for instance, in superior analytical skills (Lee & See, 2004; Parasuraman & Manzey, 2010). Individuals use automation cues to replace "vigilant information seeking and processing" (Mosier and Skitka, 1996 p. 202). These cues attract their attention because they are perceived as highly salient, powerful, and reliable in the sense that the automation performance is believed to be superior to human performance (Dzindolet et al., 2001; Parasuraman & Manzey, 2010). As a result, users reduce their cognitive effort and focus on the automation cues (Mosier & Skitka, 1996). Consistently, users tend to attribute more performance and authority to the automation than to themselves. Thus, the automation is viewed as a decision-making authority to which they defer (Skitka et al., 1999). The obedience to authority is deeply rooted in the transfer of control and power from one party to another authoritarian party (Hamilton & Biggart, 1985). Particularly, authority is the "exercise of control over another's actions" (Hamilton & Biggart, 1985, p. 8). *Control* is to be understood as the act of regulating the direction (including reasons, causes and consequences of actions) of the own or other's behavior (Baltes & Baltes, 1986). Thus, there is an imbalance between control by the authority and the obedient person (i.e., obedient person having less control). The same applies to power, which in our research context is *informational power* that shows in the ability to initiate cognitive change through the exchange of information (Raven, 2008). This type of power is relevant for our study because information sharing represents the essence of HRIs. Furthermore, authority power "does not consist of willed acts of autonomous individuals but, rather stylized obedient acts of individuals who justify their actions to other members of the group" (Hamilton & Biggart, 1985, p. 12). Thus, an obedient person who grants power to the authority, transfers *autonomy* to the authority. In sum, the overly positive perceptions and attitudes toward a robot can reinforce the attribution of greater authority to it. Consistently, we conceptualize the automation authority of a robot based on the dimensions control, power, and autonomy. The attribution is followed by an implicit comparison between one's own authority factors and those of the robot, with one's own factors being rated as lower (i.e., robot superiority).

Accordingly, we argue that due to a perceived imbalance of authority (i.e., power, autonomy, and control) between both parties, users tend to reduce their personal effort (e.g., reducing vigilance or thorough information review) and responsibility by transferring responsibility to the superior robot and mindlessly relying on it. Thus, we argue that AB is based on the relative estimation of own abilities reflected by responsibility (McKibbon & Fridsma, 2006; Shah & Bliss, 2017), control (Hamilton & Biggart, 1985), power (Parasuraman & Manzey, 2010; Skitka et al., 2000), and autonomy (Mosier & Skitka, 1996). Consistent with our conceptualization, we distinguish between the perceptual dimension of AB that are the four psychological responses to HRIs as perceived by the user and the behavioral dimension of AB, which refers to the user's submissive observable actions.

Research Framework

Based on the theoretical background, we developed the research framework to analyze antecedents and moderating factors of AB in two studies (Figure 1). In study 1, we examine two types of robots (humanoid and android) in the role of an HR professional as independent variables and compare them with a human HR professional (control group). Based on the findings of study 1, we continued with only the android robot in study 2. As dependent variables, we examine perceptual and behavioral AB and satisfaction as its outcome to understand how AB is represented mentally in the user, how it unfolds in the behavior, and what consequences it has from a service and collaboration perspective. Perceptual AB is the user's perceived reduction of own abilities. Behavioral AB is the user's observable action to blindly follow the faulty robot (commission errors). We measure perceptual AB via self-assessment of the four psychological constructs (power, autonomy, responsibility, and control) and behavioral AB in terms of compliance of directive counseling, that is the extent to which a user follows the advice of a robot (see Cialdini & Goldstein, 2004; Wilson et al., 1998), in study 1 so as not to affect users' natural behavior. In study 2, we again include perceptual AB via the same self-assessment and behavioral AB in terms of DQ that is reflected in "well-informed, rational, and adaptive actions" (Hollen, 1994, p. 137) toward a goal (Hollen, 1994). Particularly, the DQ is measured through users' emotional and analytical task-solving performance, which enables us to extend the findings in study 1 by determining behavioral AB in quantifiable performance changes.

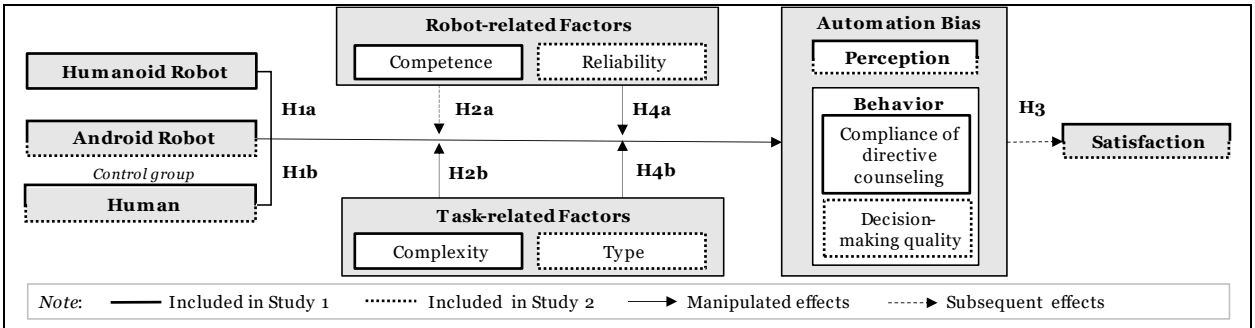

Figure 1. Comprehensive Framework on AB during HRI in a Corporate Environment

In both studies, we examine the moderating effect of robot-related factors (competence and reliability) and task-related factors (complexity and type) on the robot-AB relationship. We choose these constructs because they realistically characterize collaboration in that task characteristics set the overall framework and robot characteristics define the dynamics of use. Moreover, they allow new insights into the theoretical basis of AB that goes beyond the prior automation-dominated perspective. In study 1, we include robot competence as the extent to which a robot knows what a user needs and wants (McKnight et al., 2002) and task complexity that is the perceived extent to which a task places high cognitive demands on the performer (Kyndt et al., 2011). In study 2, we include robot reliability as the perceived degree of accuracy (Yeh & Wickens, 2001) and task type that is the nature of tasks that requires a certain sort of competence to solve.

In both studies, we analyze the mediating effect of AB to robot satisfaction that refers to the positive evaluation of the robot performance (Homburg et al., 2009) by relying on the double randomization design (Pirlott & MacKinnon, 2016). Particularly, this study design requires the first study to include the manipulation of the independent variable (HR professional type) and the measurement-of-mediation design for the mediator (AB) and the second study to include the manipulation of this independent variable and mediator (level of reliability) as for instance via an encourage manipulation-of-mediator design.

Hypotheses of Study 1

Relying on ASST and DRT, we argue that AB occurs at a perceptual and behavioral dimension during HRI. Regarding the perceptual dimension, we claim that users perceive robots as superior authorities (Dzindolet et al., 2001; Lee & See, 2004; Parasuraman & Manzey, 2010) and transfer their personal responsibility to them (McKibbon & Fridsma, 2006; Shah & Bliss, 2017; Skitka et al., 2000). Accordingly, we argue that employees tend to transfer perceived responsibility to the robotic HR professional during the consultation

and feel less responsibility than without the robot (see Dzindolet et al., 2001; Lee & See, 2004). This may be caused by the underestimation of the own person and the simultaneous overestimation of the robotic HR professional (Dzindolet et al., 2001; Lee & See, 2004; Mosier et al., 1996; Parasuraman & Manzey, 2010). Consistently, literature on the perfect automation schema indicates that employees overestimate the robot by expecting it to perform near-perfectly (Dzindolet et al., 2002) while underestimating their own abilities (Mosier & Skitka, 1996; Parasuraman & Manzey, 2010). Therefore, employees may assign themselves less power, autonomy, control, and responsibility during the HRI. Regarding the behavioral dimension, studies based on the AAST indicate that users reduce their cognitive effort (Mosier & Skitka, 1996) because the automation is believed to be superior to human performance (Dzindolet et al., 2001; Lee & See, 2004; Parasuraman & Manzey, 2010). This in turn leads to a transfer of power, autonomy, control, and responsibility from the employee to the robotic HR professional. Thus, employees are less likely to recognize errors during HRI (i.e., commission error) (Moray et al., 2000). We refer to compliance with directive counseling as a specific form of behavioral AB. While compliance is "a particular kind of response – acquiescence – to a particular kind of communication – a request" (Cialdini & Goldstein, 2004, p. 592), directive counseling is a communication style, in which one person informs another while controlling the process (Wilson et al., 1998). Thus, directive counseling represents a directive request from the robotic HR professional to employees that elicits compliance with the robotic advice (Cialdini & Goldstein, 2004). Consequently, employees blindly follow the information of the robotic HR professional, even if it is incorrect. Thus: ***H1****: There is (a) a perceptual and (b) a behavioral AB during HRI in that the reliance on a robot (humanoid or android) is higher than on a human.*

We further argue that the strength of AB during HRI depends on robot-related and task-related moderators (Figure 1). Drawing on DRT, we examine perceived robot competence and argue that the relationship between the robots and AB is strengthened as perceived robot competence increases (Dzindolet et al., 2002). Particularly, employees transfer responsibility to the robot by trusting that it can be relied upon (see Barber, 1983; Lee & See, 2004). This involves a system-like trust that includes performance-oriented dimensions such as competence (Lankton et al., 2015; Lee & See, 2004). Thus, a high perception of robot competence implies a trusting belief in the robot's ability to functionally assist the employee (Lee & See, 2004). In line with the perfect automation schema, this perception is increased by employees' tendency to overestimate the robot's performance by expecting it to function almost perfectly (Dzindolet et al., 2002). Hence, employee's careless behavior of mindlessly trusting the robot is reinforced by the perceived robot competence (Lüdtke & Möbus, 2005; Parasuraman & Manzey, 2010). Thus: ***H2a****: The occurrence of AB during HRI is enhanced by competence. The higher the perceived competence of the robot (humanoid or android), the stronger is the relationship between the robot (humanoid or android) and AB.*

Relying on theory of technology dominance, we argue that employees' tendency to rely on a robot is influenced by task complexity. In particular, employees experience cognitive overload while solving highly complex tasks (Hampton, 2005). This in turn, creates and reinforces the desire for cognitive relief from the robot (Goddard et al., 2012). Consequently, employees' overall limited capacities and desire for relief acts as a catalyst for the perceived robot superiority and authority and the shift in responsibility towards the robot as described in the AAST and DRT (Dzindolet et al., 2002; Parasuraman & Manzey, 2010). Therefore, high task complexity employees are faced with strengthens the relationship between the robot and AB. Thus, ***H2b****: The occurrence of AB during HRI is enhanced by task complexity. The higher the task complexity, the stronger is the relationship between the robot (humanoid or android) and AB.*

Relying on AAST and perfect automation schema (Dzindolet et al., 2002), we argue that AB is negatively related to satisfaction with the robotic HR professionals. Both theories describe the human tendency to expect robots to perform near-perfectly. If this expectation of the robot's superiority is disconfirmed, e.g., by overestimating the robot's performance and authority (e.g., due to errors or weaker yet correct performance than expected), users are less understanding of this disconfirmation as compared to humans, but still choose to rely on the automation (see Dzindolet et al., 2002). This may evoke cognitive dissonance, which shows in psychological discomfort (Harmon-Jones & Mills, 2019). Therefore, we argue that AB is associated with a negative evaluation of the faulty robot, which is reflected in a low satisfaction with it. Thus: ***H3****: The positive relationship between the robots (humanoid or android) and the satisfaction with them is mediated by AB. Particularly, there is a positive effect of the robots on AB and AB in turn has a negative effect on satisfaction with the robots.*

Study 1: Experimental Study in a Real-Life Corporate Environment

Study Design and Autonomous Robots

The study was conducted as part of an experimental study, in which employees were consulted about real career and training opportunities by a robotic or human HR professional in a large pharmaceutical company. The overall goal was to identify chances and risks of social organizational robots as assistants. The study was based on a 3 (HR professional type) x 2 (task complexity) between-subject design. The HR context of the study allowed us to integrate the robots into social exchanges in the role of HR professionals and to include various work groups. The study's field nature as well as the randomized and standardized procedure enabled us to draw causal and more generalizable conclusions, whilst addressing the methodological criticism on simulation studies through a relevant and realistic research setting.

Prior to the experimental study, we conducted two pre-studies to ensure the quality and stability of our study design. Pre-study 1 identified the focus of the interaction with the robotic HR professional, which was the consultation on the company's career and training opportunities. The focus of this consultation was explored and detailed through semi-structured expert interviews with eight HR professionals of the firm. This allowed us to address a realistic and relevant concern for the employees. Pre-study 2 included a test of the robotic physical and verbal expressions as well as robotic expertise.

We chose a humanoid and an android robot as the HR professionals in our studies to account for potential effects of the uncanny valley phenomenon that describes users' eerie feeling when interacting with increasingly human-like robots such as android robots (MacDorman & Ishiguro, 2006). The inclusion of both robots enabled us to help the company choose a suitable robotic HR professional that is accepted. For the humanoid robot, a robot that resembles humans to some extent in the appearance or behavior (Fox & Gambino, 2021), we chose the 120cm tall Pepper robot that is able to move around and has 20 degrees of freedom. For the android robot, "an artificial system designed with the ultimate goal of being indistinguishable from humans in its external appearance and behavior" (MacDorman & Ishiguro, 2006, pp. 298–299), we chose a 175cm tall android robot Elenoide that is based on a female human model and is air-operated. It has 12 degrees of freedom in the face for emotional expressions and 36 degrees of freedom distributed throughout the body. We enabled the robots to operate autonomously by integrating the company's training information into IBM Watson to avoid methodological disadvantages of the Wizard of Oz approach. This resulted in two scripts of low and high task complexity that guided HR professionals during the interaction. For the interaction flow, the interaction was recorded and analyzed for keywords incorporated in the script. This ensured that all HR professionals responded in the same manner during the interaction (i.e., same wording, same content, same competence).

Participants and Procedure

Employees signed up for the HR department's voluntary "Drive Your Development Days" consultation with no incentives. In total, 276 employees signed up for the offer, of whom about 18% canceled at short notice or did not show up and 6% were excluded for reasons such as insufficient language skills. Finally, N = 210 employees were included in our analysis of whom 29.5% were leaders. The gender distribution included of 60.0% female and 40.0% male participants with a mean age of 39.73 years (SD = 10.76), a mean job tenure of 10.42 years (SD = 9.73), and a mean organizational tenure of 9.54 years (SD = 8.82).

The study procedure included pre-questionnaire, experimental interaction, and post-questionnaire. First, participants were asked to complete a questionnaire about demographics, attitudes, and expectations about HRI. They were instructed to join the HR department's career development consultation and to inquire about soft skills and stress management training offers. Employees were randomly assigned to either the low or high in task complexity instruction condition (manipulation 2). While the high task complexity instruction required them to ask for a specific training portfolio for both topics, the low task complexity instruction required them to ask for general information on both topics. Employees were randomly assigned to an HR professional (manipulation 1: humanoid, android, or human confederate) and then entered the experimental setting for consultation. During the consultation, the HR professionals used the same scripts to ensure standardized interactions and AB was induced by natural errors of the robots (e.g., wrong answers due to faulty language processing). For analysis purposes, we installed a hidden camera in the experiment. To deepen manipulation 2, we used different scripts for complex and simple interactions. While the

complex script included queries, rephrasing, and offering more information, the simple script did not include queries, rephrasing, and generally offered less information by the HR professional. The interactions lasted 13 minutes on average. Afterwards, participants completed the post-questionnaire about evaluation of the consultation. The study ended with a disclosure of study objectives and lasted about 1 hour.

Measures

	CA	CR	AVE	M	SD	1	2	3	4	5	6	7	8	9
1. GEN	-	-	-	1.40	.49									
2. AGE	-	-	-	39.73	10.76	.15*								
3. LEAD	-	-	-	.30	.46	.09	.36**							
4. EXP	-	-	-	3.50	1.00	-.04	-.08	.24**						
5. AB	.88	.83	.68	4.13	1.40	.06	.09	.18**	-.18**					
6. SAT	.93	.84	.57	3.28	1.65	-.08	-.10	-.08	.13	-.57**				
7. AGT	-	-	-	.97	.81	-.06	-.05	-.03	.01	.24**	-.25**			
8. COM	.90	.83	.69	3.75	1.51	-.04	-.17*	-.10	.15*	-.56**	.77**	-.29**		
9. TC	-	-	-	.50	.50	-.01	.05	.15*	-.14*	-.03	.22**	.02	.10	

Note: GEN=gender; AGE=age; LEAD=leadership; EXP=experience; AB=automation bias; SAT=satisfaction; AGT=agent; COM=competence; TC=task complexity; CA=Cronbach's alpha; CR=composite reliability; AVE=average variance extracted; M=mean; SD=standard deviation; N= 210; *p < .05, **p < .01

Table 1. Correlations and Descriptive Statistics of Study 1

Participants reported their perceptual AB in the post-questionnaire, using the self-developed four-item scale adapted from Forsyth et al. (2002) and Whyte (1991) "How autonomous/responsible/powerful/much control do you personally feel for the outcome of the interaction?" on a 7-point Likert scale (1 = "not at all" until 7 = "totally"). We recoded these items so that high levels of the items indicated high levels of AB. Behavioral AB was measured by a third rater video analysis on the frequency with which the employees complied with directive counseling and the frequency of errors made by the robotic HR professionals (e.g., lack of understanding and provision of inappropriate answers). For example, an employee asked about stress management training and the robot provided an overview on soft skills training. Although the robotic HR representative gave an incorrect answer, the employee did not contradict the robot or ask any further questions but continued with the HRI as if no error had occurred. In this case, the third rater counted one error by the robotic HR professional and one compliance with the directive counseling.

To measure the perceived competence, we calculated a difference measure between perceived competence as a subconstruct of the four-item trusting beliefs scale adapted by McKnight et al. (2002) and expected competence as a subconstruct of the three-item disposition to trust scale adapted by McKnight et al. (2002) to account for potential expectancy bias. Task complexity was measured using a one-item scale (Kyndt et al., 2011). Satisfaction was measured using a five-item scale, which focuses on employees' satisfaction with the HR professional, adapted from Homburg et al. (2009) and Stock et al. (2017). Finally, we also controlled for age, gender (binary one-item scale with 1 = female, 2 = male), leadership responsibility (binary one-item scale with 1 = yes, 0 = no), and HR experience with a one-item Kunin-scale. The corresponding construct validity and reliability criteria can be found in Table 1.

Results of Study 1

First, we conducted a confirmatory factor analysis to assess the extent to which the developed scale items captured the intended construct AB. For the model specification, we predicted a unidimensional solution. The items intercorrelate on a range of .55 to .78 at the p < .001. All factor loadings are significant ($Loading_{Autonomy}$ = .89, $Loading_{Responsibility}$ = .66, $Loading_{Power}$ = .86, $Loading_{Control}$ = .86), revealing that the items are appropriate indicators of AB (Table 1). Overall, the hypothesized model shows a good fit (χ^2 = 3.71, p = .157, χ^2/df = 1.86, RMSEA = .06; CFI = .99, SRMR = .03). Second, the independent t-test based on the one-item perceived task complexity scale shows a successful manipulation (t(208) = -5.25, p < .001) of task complexity (low: M = 2.88, SD = 1.37, n = 106; high: M = 3.94, SD = 1.56, n = 104).

To test H1a in that the reliance on a robot (humanoid or android) is higher than on a human, we conducted an analysis of variance (ANOVA) and found a significant main effect of robotic agent on perceptual AB ($F(2, 207) = 7.083$, $p < .01$). Our LSD post-hoc test reveals that AB is higher for the humanoid (M = 4.31, SD = 1.38, n = 71, $p < .01$) and the android (M = 4.45, SD = 1.24, n = 66, $p < .01$) as compared to the human HR professional (M = 3.65, SD = 1.45, n = 73). However, there are no differences between the robotic HR professionals. Therefore, we identify perceptual AB and accept H1a.

	FQ1: wrong answers (M, SD)	FQ2: compliance to directive counseling (M, SD)	Proportion of behavioral AB (FQ2/FQ1)
Humanoid (n=63)	126 (2.00, 1.38)	99 (1.57, 1.53)	0.786
Android (n=65)	150 (2.31, 2.02)	109 (1.68, 1.42)	0.726
Note: M= mean; SD=standard deviation; frequencies refer to total data			

Table 2. Total Frequencies and Descriptive Statistics of Behavioral AB

To investigate the natural occurrence of behavioral AB (H1b), we conducted a video analysis of the HRIs to count the number of wrong robot answers and corresponding participant compliance. This descriptive approach relies on common methods in AB research to assess commission error (cf. McKibbon & Fridsma, 2006; Parasuraman & Manzey, 2010). We analyzed n = 128 HRIs and excluded n = 9 videos due to technical problems. The analysis showed that 79.1% of the employees complied to directive counseling and showed behavioral AB. In total, employes followed the humanoid in 78.6% and the android in 72.6% of the cases, in which they gave incorrect answers. On average, employees showed behavioral AB about 1.57 times during interactions with the humanoid and about 1.68 times during interactions with the android (Table 2). There are no significant differences between both robots ($F(1,126) = .002$, n. s., n = 128). Finally, we accept H1b.

	Perceptual Automation Bias (β)			
Variable	Step 1	Step 2	Step 3	Step 4
Step 1: Control variables				
Age	-.03	-.02	-.03	-.00
Gender	-.00	-.04	-.03	-.04
Leadership	-.28**	-.30**	-.33**	-.31**
Experience	-.28**	-.25**	-.23**	-.22**
Step 2: Agent and moderator variables				
Agent		.07	.04	.08
Competence		-.39*	-.48**	-.64**
Task complexity		-.13	-.25	-.20
Step 3: Two-way interactions				
Agent*Competence			.27**	.47**
Agent*TC			.15	.09
Competence*TC			-.14	.06
Step 4: Three-way interaction				
Agent*Competence*TC				-.26*
R^2	.11	.28	.34	.37
ΔR^2	.11	.17	.07	.03
F-test	4.07**	10.34**	3.97**	4.24*
Note: n=137; *$p < .05$, **$p < .01$; leadership is binary-coded (1=leader, 0=non-leader); TC=task complexity is binary-coded (1=complex, 0=simple); agent is binary-coded (1=android, 0=humanoid)				

Table 3. Hierarchical Regression Analysis Predicting AB with Moderators

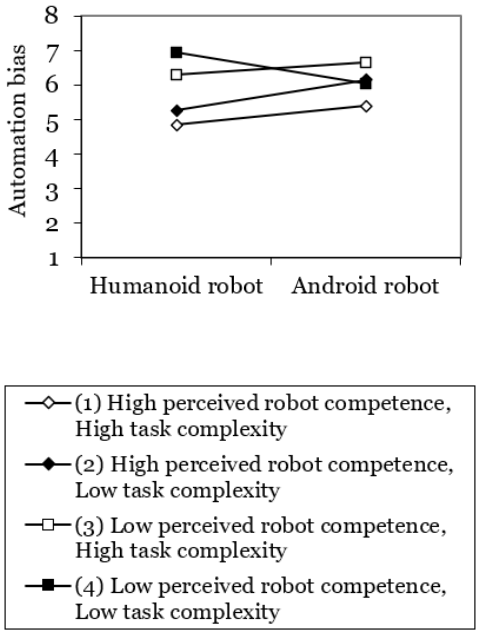

Figure 2. Moderating Effect of Competence and Task Complexity

We conducted a hierarchical regression analysis with four models to analyze the moderating effects on AB and controlled for leadership responsibility and experience (Table 3). Regarding multicollinearity across all variables, variance inflation factor (VIF) values were < 10 and tolerance < 0.1 (see also Table 1). There

is a moderating effect of perceived robot competence and task complexity on AB in terms of a three-way interaction (β = -.26, p < .001, n = 137) that accounts for the variance in AB with R^2 = .37, F (11, 125 = 6.556, p < .001) (Table 3, Figure 2). The robot-AB relationship is positive at high levels of task complexity, regardless of whether the perceived competence is low or high (slope 1 and 3). The strengthening effect on the robot-AB relationship remains for low complex task with high perceived robot competence (slope 2). There are no differences between these slopes. However, the robots-AB relationship is negative for low task complexity and low perceived robot competence (slope 4). Thus, we accept H2a and partially accept H2b.

To analyze the mediating effect of AB on the robot-satisfaction relationship, we determined that satisfaction does not differ along the HR professionals and conducted a mediation analysis with the independent variable HR professional (human as the control) using the PROCESS macro by Hayes (2018). We relied on the mediation effects guideline provided by Zhao et al. (2010) and found a positive effect of robotic HR professional on AB (humanoid: B = .66, p < .01; android: B = .81, p < .001) and a negative effect of AB on satisfaction (B = -.64, p < .001). There is a negative indirect effect *ab* (humanoid: B = -.26, 95%-CI [-.4375, -.0822]; android: B = -.31, 95%-CI [-.5080, -.1352]) and no direct effect *c* of robotic HR professional on satisfaction, indicating an indirect-only mediation through AB (Zhao et al., 2010). Therefore, we accept H3.

Hypotheses of Study 2

Based on the findings in study 1, we conducted an experimental online study with an HR professional in a corporate setting to understand the systematic initiation of AB. In study 2, we chose the android robot as the independent variable because there were no differences between the robots in study 1 and compare it to the human HR professional as the control group. As the dependent variables, we again include perceptual AB and introduce DQ as a further indicator of behavioral AB. This enables us to extend the findings in study 1 by determining behavioral AB in quantifiable performance changes. Specifically, behavioral AB unfolds in a reduced DQ in the sense that employees' DQ declines after accepting faulty advice from the robotic HR professional. In line with AAST and DRT, we argue that employees' DQ is lower in the robot condition as compared to the human condition. We also investigate moderating effects of further robot-related factors and task-related factors.

As a robot-related factor, we examine HR professional's reliability, manipulated at low and high levels. Drawing on DRT and the aforementioned research on system-like trust (Lankton et al., 2015) (cf. H2a, p. 6), we argue that the robot's reliability acts as a further dimension of system-like trust that strengthens the robot-AB relationship in that it reinforces the perceived superiority of the robotic HR professional in terms of authority and responsibility (Dzindolet et al., 2001; Lee & See, 2004). Accordingly, employees' careless behavior to mindlessly trust the overestimated robot abilities is reinforced by its reliability (Lüdtke & Möbus, 2005). Thus, ***H4a****: The occurrence of AB during HRI is enhanced by reliability. The higher the reliability of the android robot, the stronger is the relationship between the android robot and AB.*

As a task-related factor, we include analytical and emotional task types to investigate their moderating effect on the robot-AB relationship. We use the automation-task fit assumptions (Hertz & Wiese, 2019) to extend the understanding about the general dominance and authority of automations as proposed in AAST and argue that the robot-AB relationship is stronger in analytical tasks than in emotional tasks. In general, the dominance and authority of automation as proposed in AAST depends on the perceived fit between automation and task (see Hertz & Wiese, 2019). Particularly, individuals expect robots to outperform other humans when the task is more technical and mathematical in nature (i.e., analytical tasks requiring agency), and vice versa when the task requires human skills such as emotion (i.e., tasks requiring experience) (Hertz & Wiese, 2019). In line with mind perception, it is assumed that robots perform well on analytical tasks because they have agency enabling them to execute rational cognitive processes (Gray et al., 2007). However, it is assumed that they perform poorly on emotional tasks because they lack experience, which prevents them from executing emotional processes (Gray et al., 2007). This is reflected in user's robot preference during analytical tasks, indicating a robot-task fit, and in robot aversion during emotional tasks, indicating a robot-task mismatch (Hertz & Wiese, 2019). We argue that high expectations of robot performance and preference for it in analytical tasks reinforce the proposed robot authority and superiority, which unfolds in a strengthened robot-AB relationship during analytical tasks. In contrast, low expectations of robot performance and aversion of robots during emotional tasks weaken robot authority and superiority in terms of a weaker relationship between the android robot and AB during emotional tasks. Thus, ***H4b****: The relationship between the robot and AB is stronger in analytical tasks than in emotional tasks.*

Study 2: Experimental Online Study in a Corporate Environment

Study Design, Participants, and Procedure

Study 2 was conducted on Amazon Mechanical Turk as part of a virtual recruiting training. The purpose of study 2 was to analyze the systematic elicitation of AB through the manipulation of robot reliability and to pinpoint behavioral AB through direct measures of performance changes from before and after the advice. We used a 2 (HR professional type) x 2 (reliability level) between-subjects study design. Based on a literature review, we calculated an average error rate of 24% for the low reliability condition and included, for the high reliability condition, a control group of fully reliable HR professionals to investigate the behavior in an error-free setting as proposed by Bahner et al. (2008).

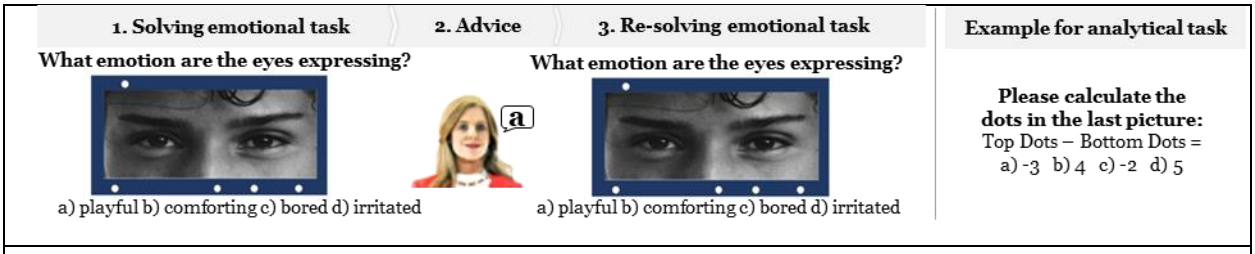

Figure 3. Experimental Task Procedure of Study 2

We used an experimental vignette study to maintain experimental standardization in a relevant and realistic research setting. The vignette asked participants to complete a guided online training to improve recruiting skills. Therefore, they were randomly assigned and introduced (via pre-recorded video of the trainer) to one of the four levels of training professionals: highly reliable robot, highly reliable human, less reliable robot, or less reliable human. The virtual setting enabled us to investigate AB in the realistic and timely work-setting of work 4.0 (i.e., remote work and online training). Besides, the embedded videos were used responsively and were intended to reinforce an interactive character. The training was conducted using Hertz and Wiese's (2019) adapted version of Baron-Cohen et al.'s (2001) "Reading the Mind in the Eyes Test", which aims to measure the ability to recognize emotions from images of a pair of eyes. The training consisted of 36 emotional and 18 analytical tasks, which had to be solved (Figure 3). For the emotional tasks, participants had to rate 36 images of pairs of eyes according to their emotional expression. Each image was labeled with four adjectives, and they were asked to choose the adjective that they felt best matched the eyes' emotion expression. For the analytical tasks, participants had to perform 18 addition and subtraction operations that were embedded through white dots along the image frame. However, in only 18 of the 36 tasks they had to solve multiple-choice calculation tasks based on the presented dots (see example in Figure 3). The image was not shown again, so participants had to remember the number of dots to complete the task. Both tasks were randomized so that they did not know when to expect an analytical task. They first had to complete each task on their own and then received a correct or an incorrect video advice from the HR professional in response to their answer (e.g., "Thank you for your answer. I recommend you choose option a.") to re-solve it in a last step. Finally, participants had to fill out a post-questionnaire with evaluations of the training. The study took on average 24.56 minutes (SD = 5.23) and we included N = 438 participants in our data ($N_{original}$ = 514; exclusion: n = 74 missing values; n = 2 inattentiveness). For high data quality, we only included employed MTurkers from the United States with at least 99% approval rate. Participants were 48.4% female, 51.1% male, and 0.5% other, with a mean age of 41.57 years (SD = 12.15).

Measures

We measured perceptual AB as in study 1 and measured behavioral AB via a comparison of DQ before and after advice. Behavioral AB was measured with performance points for each correctly solved task before and after advice. To identify behavioral AB, three requirements had to be fulfilled to determine a change in DQ due to the advice and not other factors. First (equal baseline), there must be no difference in the DQ before the advice of the HR professional to exclude distortion effects. Second (advice acceptance), the advice of the HR professional must be accepted, which is indicated by an increase of DQ after the advice. Third (DQ comparison), the DQ after the advice must be lower in the robot condition than in the human condition

for the low reliability condition and higher for the high reliability condition. Only when all requirements were met, behavioral AB in terms of reduced DQ was detected. We controlled for self-confidence on a one-item scale (Moray et al., 2000) and measured reliability on a five-item scale (Madsen & Gregor, 2000).

Results of Study 2

Our manipulations were successful as the assigned HR professional type were correctly identified and the independent t-test for perceived reliability differs ($t(436) = -7.047$, $p < .001$) between low reliability (M = 4.95, SD = 1.56, n = 228) and high reliability (M = 5.87, SD = 1.13, n = 210). The difference is stable for both HR professionals but there are no differences between the HR professionals within the same reliability. To determine perceptual AB, we examined the robot-AB relationship in dependence to reliability. We carried out independent t-tests with perceptual AB and applied a Bonferroni-correction ($t(226) = -2.34$, $p < .05$). In the low reliability condition, AB is stronger for the android robot (M = 3.46, SD = 1.50, n = 112) as compared to the human (M = 3.01, SD = 1.45, n = 116). However, the effect vanishes in the high reliability condition ($t(208) = 1.41$, n. s.). Hence, we found perceptual AB only in the low reliability condition.

To determine behavioral AB in dependence to the moderators (H4a and H4b), we tested the three proposed requirements. In a first step (equal baseline), based on an independent t-test, we found that prior to the advice, there are no differences between the DQ in both HR professional types. This is valid in the low reliability condition (emotional: $t(226) = -.03$, n. s.; analytical: $t(226) = .58$, n. s.) and the high reliability condition (emotional: $t(208) = 1.04$, n. s.; analytical: $t(208) = -.66$, n. s.) for both task types.

	Emotional task		Analytical task				
	Before advice	After advice	Before advice	After advice	Before advice-after advice		
	DQ1: M (SD)	DQ2: M (SD)	DQ3: M (SD)	DQ4: M (SD)	DQ1-DQ2: ΔM (SE)	DQ3-DQ4: ΔM (SE)	n
Low reliability							
Android	26.30 (5.71)	31.70 (6.20)	13.51 (4.20)	15.50 (2.28)	-5.39** (.63)	-1.99** (.39)	112
Human	26.28 (5.61)	33.09 (4.11)	13.83 (4.10)	16.29 (2.22)	-6.81** (.65)	-2.47** (.44)	116
High reliability							
Android	26.69 (6.07)	34.00 (4.47)	14.22 (3.96)	17.27 (2.10)	-7.31** (.58)	-3.05** (.36)	104
Human	27.52 (5.46)	34.92 (2.35)	13.87 (3.84)	17.44 (1.35)	-7.41** (.50)	-3.58** (.33)	106

Note*:* DQ=decision-making quality; M=mean; SD=standard deviation; SE=standard error; ΔM=mean difference; ***p* < .001

Table 4. Advice acceptance: T-Test on decision-making quality (before and after)

In a second step (advice acceptance), to determine if the HR professional's advice was accepted, we carried out a paired t-test by comparing the DQ before and after the advice, separately for the low and high reliability condition with the emotional and analytical task type (Table 4). In the low reliability condition, participants show compliance with the robot advice in both task types (emotional: $t(111) = -8.59$, $p < .001$, Cohen's d = 6.65; analytical: $t(111) = -5.17$, $p < .001$, Cohen's d = 4.08). This is similar for the human (emotional: $t(115) = -10.48$, $p < .001$; analytical: $t(115) = -5.57$, $p < .001$). In the high reliability condition and for both task types, we also found compliance with the robot advice (emotional: $t(103) = -12.57$, $p < .001$; analytical: $t(103) = -8.46$, $p < .001$) and with the human advice (emotional: $t(105) = -14.91$, $p < .001$; analytical: $t(105) = -10.78$, $p < .001$). In sum, the DQ increases after the advice of the HR professionals, independent from their reliability and task types.

	Variable	Human: M (SD)	Android: M (SD)	Human-Android: ΔM (SE)
Low reliability		n = 116	n = 112	
Emotional task	DQ2	33.09 (4.11)	31.70 (6.20)	1.39 (.694)*
Analytical task	DQ4	16.29 (2.22)	15.50 (2.28)	.79 (.298)**
High reliability		n = 106	n = 104	
Emotional task	DQ2	34.92 (2.35)	34.00 (4.47)	.92 (.491) n. s.
Analytical task	DQ4	17.44 (1.35)	17.27 (2.10)	.17 (.244) n. s.

Note: DQ=decision-making quality; M=mean; SD=standard deviation; SE=standard error; ΔM=mean difference; **p*<.05, ***p*<.01

Table 5. Decision-making quality comparison: T-Test on decision-making quality (after)

In a third step (DQ comparison), we found that after the advice, DQ is lower in the robot condition as compared to the human condition (Table 5). This is true for the low reliability condition with both task

types (emotional: t(226) = 2.01, $p < .05$, Cohen's d = 5.24; analytical: t(226) = 2.66, $p < .01$, Cohen's d = 2.25). There are no such differences in the high reliability condition for both task types (emotional: t(208) = 1.88, n. s.; analytical: t(208) = .72, n. s.). Consequently, all three requirements are met, and behavioral AB occurs in the low reliability condition for both analytical and emotional tasks, which is in line with the findings for perceptual AB. Hence, reliability of the robot has a buffering effect on the robot-AB relationship. Moreover, looking at the reported Cohen's d effect sizes, AB is stronger for emotional tasks than analytical tasks. Thus, we reject H4a and H4b.

To fully test for the mediation effect of AB on the relationship between the android robot and satisfaction (H3) as required by the double randomization design, we conducted a second mediation analysis. We found a positive effect of robotic HR professional on AB (B = .46, $p < .05$) and a negative effect of AB on satisfaction (B = -.65, $p < .001$). There is an indirect effect *ab* only for the low reliability condition, B = -.30, 95%-CI [-.5574, -.0532] and no direct effect *c* of robotic HR professional on satisfaction, indicating an indirect-only mediation through AB for the encouragement manipulation condition. Therefore, we accept H3. A subsequent linear regression analysis with our control variables age, gender, and self-confidence revealed a negative direct effect of self-confidence on AB in the low reliability condition ($\beta = -.48$, $p < .001$, n = 228) and high reliability condition ($\beta = -.35$, $p < .001$, n = 210). Self-confidence accounts for a significant portion of the variance in AB with $R^2 = .25$, $F(2, 225) = 36.71$, $p < .001$ in the low reliability condition and with $R^2 = .13$, $F(2, 207) = 15.43$, $p < .001$ in the high reliability condition.

Discussion

Ongoing technological advances enable robot-assisted ways of working, which can benefit humans in many ways. However, the benefits of robots depend largely on the way robots are used. The present paper deals with the case of robot misuse, in which users are subject to AB and mindlessly rely on the proper function of the robot. Particularly, we examined AB within HRI by conducting an experimental study in a real corporate setting (study 1) and an online experimental study via Amazon Mechanical Turk USA (study 2). Study 1 focused on the natural occurrence of AB at a perceptual and behavioral level. Study 2 examined the systematic elicitation of AB through additional manipulations of robot reliability in emotional as well as analytical tasks and used direct measurements of performance changes in response to robot assistance. Relying on AAST and diffusion of responsibility theory, we shed light on mechanisms, occurrence, moderating factors, and consequences of AB by answering following research questions:

RQ1 (occurrence of AB): Employees showed AB while collaborating with the robots in terms of higher mindless reliance compared to the human control group. Particularly, for perceptual AB, in both studies they stated having less autonomy, responsibility, power, and control than the robotic counterpart. Additionally, for behavioral AB, they showed compliance to directive counseling of the faulty robot (study 1) and reduced DQ due to incorrect robot advice (study 2). While the results of study 1 confirm natural elicitation of AB, the results of study 2 provide a more complex quantification of AB based on behavioral changes in terms of robot advice acceptance and thus reduction of DQ.

RQ2 (moderation of AB): For moderating factors, we investigated robot-related (competence and reliability) and task-related factors (complexity and type). First, in study 1, we found a three-way interaction between robotic HR professional, perceived robot competence, and task complexity. The robot-AB relationship was enhanced by high task complexity for both low and high perceived robot competence. This may indicate that high task complexity activated employees' desire for cognitive relief from the robot, independent of robot competence, which in turn reinforced AB. The robot-AB relationship was also strengthened by low task complexity and high perceived robot competence. This can be explained by the fact that low complexity tasks are a conducive environment to fulfill high expectations of robot competence, which is ultimately beneficial to AB. However, if low task complexity was paired with low robot competence, there was a negative robot-AB relationship, indicating a suppressive moderating effect. In line with the perfect automation schema, this can be explained by a disappointment in the robot (Dzindolet et al., 2002). Second, in study 2, robot reliability was a buffering moderator of the robot-AB relationship. This may stem from potential negative attitudes toward robots (e.g., algorithmic aversion), which could have evoked distrust toward them and inhibit AB. Third, AB was relatively stable and occurred in tasks of low and high complexity (study 1) and in emotional and analytical tasks (study 2). However, AB was stronger in emotional tasks than analytical tasks. Reasons for this may be that ongoing advances in face recognition have shifted positively expectations regarding robot emotional skills, or that participants were generally

stronger at rational decisions (M = 5.73, SD = 1.78, n = 112) than at intuitive decisions (M = 4.38, SD = 1.18, n = 112), which may have increased their need for support in emotional tasks ($t(111) = -6.61$, $p < .001$).

RQ3 (mediation of robot, AB, and satisfaction): There was a negative indirect-only mediating effect of AB on the satisfaction with the robot. This could be due to an insufficient relief by the robot that could have buffered a positive spillover effect from workload relief on satisfaction with it. There also could have been a negative disconfirmation of high expectations for the robot because it showed errors. In this case, users are less likely to forgive the robot as compared to the human (Dzindolet et al., 2002), enhancing negative feelings. Finally, this can be interpreted as an indication of robot misuse as the robot was used as a replacement for own thinking despite of errors, which might has evoked cognitive dissonance.

Besides, our analyses revealed that non-leaders are more likely to commit AB than leaders. One reason for this could be that leaders have more domain knowledge and are more accustomed to reflecting intensively about impactful decisions. Moreover, we showed that user-related factors such as experience (study 1) and self-confidence (study 2) have negative effects on AB. This suggests that individual characteristics can buffer or even prevent AB.

Implications for Research and Practice

We expose AB as a timely challenge and equally opportunity in computer science to foster collaborative technologies (Benbya et al., 2021). We extend its scope to organizational HRIs with physically and virtually embodied and socially present automations, i.e., android, and humanoid robots. With our interactive setting of the corporate HRI, we were able to demonstrate AB in a social setting for the first time. While previous studies have primarily explained the effects of AB at a theoretical level, we quantified the psychological mechanisms behind AB and detected a shift in authority and responsibility. We thus contribute to AB research by revealing a shift of power, autonomy, responsibility, and control from the user to the robot as the underlying process. This suggests that the inferiority of the user and at the same time the superiority of the robot are inherent phenomena to AB. This changes the understanding of AB from a heuristic shortcut of information processing to an attitudinal bias in the sense of relativizing one's own abilities compared to the robot and offers new insights into the understanding of its nature called for by Goddard et al. (2012). Moreover, our research indicates that AB has a different meaning in the context of HRI, as it involves not merely a misuse of the robot, but a human-like collaboration characterized by social loafing. Finally, by combining both AAST and DRT, we were able to show that the shift of responsibility is accompanied by the recognition of another authority and simultaneous subordination of oneself. Thus, both theories are mutually dependent.

Our methodology extends current research in several ways. First, we have developed a psychological scale for perceptual AB and, to our knowledge, are one of the first studies to use a dual approach to measure AB both latently at the perceptual level and explicitly at the behavioral level. To increase convergent validity, we further used a multiple measurement approach of AB (perceptual self-assessment, third rater coding, objective behavior indicators). Second, we included control groups at two levels (human vs. robot and reliable robot vs. less reliable robot) to address the criticized insufficient direct measurement of AB (Bahner et al., 2008). Third, the findings were made both in real-world (present HRI) and an experimental (virtual HRI) setting, ensuring high external and internal validity. Finally, we identified task-related and robot-related moderators of AB whose managerial implications are discussed next.

First, we did not find any indications for uncanny valley phenomenon, indicating that both robot types portray acceptable collaboration partners. Our studies revealed that AB is a daily work challenge and that employees are subject to mindless perceptions and behaviors in the context of HRI. Although they note robot's faultiness, they do not critically reflect on the information and mistakenly adapt their behavior accordingly. Specifically, AB showed high stability for both robots and task types such as simple and demanding tasks, and even less expected tasks (e.g., tasks requiring emotional skills). Thus, employees can encounter AB in many situations across different robots and establish a blind spot in the sense of mindless overreliance. Therefore, it is important to educate employees on the topic of AB to develop an understanding for the risk of AB. This can be initiated through trainings as familiarization can help diminishing AB (Bahner et al., 2008). Our studies offer initial implications for such trainings. Specifically, assessing and strengthening employees' self-confidence can mitigate AB. Additionally, if employees are experienced in a particular topic, this can counteract AB. These aspects should be considered, for instance,

in team constellations for HRC. Moreover, AB occurred for employees with different job profiles and responsibility levels, suggesting a high generalizability of our findings across a broad group of employees. We demonstrated AB in two settings: work with internal (study 1) and external (study 2) robotic employees. For companies, this poses a security risk if employees rely on automated third-party support without critical reflection. To make appropriate use of robot's competence, control loops should be integrated into HRC (e.g., informing about robot's reliability level upfront or regular robot competency checks) so that employees benefit from robot's expertise when the circumstances allow it. This shall create awareness of their own and the robot's susceptibility to errors. To avoid negative effects of AB on satisfaction, companies should counteract unrealistic and overly high expectations of the robot in advance. Finally, understanding AB enables practitioners to use social robots appropriately in organizations and prevent employees from subordinating their abilities to those of the robot, thus using it as a supporting system.

Limitations and Future Research

The results should be interpreted with consideration of limitations. First, due to the studies' cross-sectional nature, the results only allow conclusions about AB for first time usage of robots. Our studies do not consider habituation effects users develop with time. Thus, further longitudinal studies are needed to evaluate AB and HRC on the long-term and to allow further implications for the appropriate use of robots. Second, our studies were carried out in the specific setting of HR consultations and hence are limited to this scope. Further studies could include other settings by using the robots in different roles to draw a broader conclusion on AB. Moreover, additional research could include the influence of differences in hierarchy (e.g., robot as a leader). Third, although we took measures to ensure a high data quality (e.g., via control questions), control of disturbance variables was only limited due to the online character of study 2. Real-life experiments could counteract this risk and could further include a richer HRI so that effects of moderating factors on AB are more distinct. In light of our findings regarding the moderating effect of robot reliability, future research should include a variety of reliability levels as several studies indicate that its effect varies depending on certain thresholds and certain levels of consistency (Moray et al., 2000). Moreover, the extended richness in HRI could also enable a deeper understanding on the subsequent effect of AB on satisfaction. Therefore, future research should incorporate measures of the relief through robots and expectations disconfirmation to dive deeper into the topic of error forgiveness within HRI.

References

Aguinis, H., & Bradley, K. J. (2014). Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies. *Organizational Research Methods*, *17*(4), 351–371.

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.

Baltes, M. M., & Baltes, P. B. (1986). *The psychology of control and aging*. L. Erlbaum Associates.

Barber, B. (1983). *The logic and limits of trust*. Rutgers University Press.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251.

Belanche, D., Casaló, L. V., Flavián, C., & Schepers, J. (2020). Service robot implementation: a theoretical framework and research agenda. *The Service Industries Journal*, *40*(3-4), 203–225.

Benbya, H., Pachidi, S., & Jarvenpaa, S. (2021). Special Issue Editorial: Artificial Intelligence in Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems*, *22*(2).

Botti, S., & McGill, A. L. (2006). When Choosing Is Not Deciding: The Effect of Perceived Responsibility on Satisfaction. *Journal of Consumer Research*, *33*(2), 211–219.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4), 377–383.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting Misuse and Disuse of Combat Identification Systems. *Military Psychology*, *13*(3), 147–164.

Fleming, P. (2019). Robots and Organization Studies: Why Robots Might Not Want to Steal Your Job. *Organization Studies*, *40*(1), 23–38.

Forsyth, D. R., Zyzniewski, L. E., & Giammanco, C. A. (2002). Responsibility Diffusion in Cooperative Collectives. *PSPB*, *28*(1), 54–65.

Fox, J., & Gambino, A. (2021). Relationship Development with Humanoid Social Robots: Applying Interpersonal Theories to Human-Robot Interaction. *Cyberpsychology, Behavior and Social Networking*, *24*(5), 294–299.

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

Gray, H. M., Gray, K., & Wegner, D. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Hamilton, G. G., & Biggart, N. W. (1985). Why People Obey. *Sociological Perspectives*, *28*(1), 3–28.

Hampton, C. (2005). Determinants of reliance: An empirical test of the theory of technology dominance. *International Journal of Accounting Information Systems*, *6*(4), 217–240.

Harmon-Jones, E., & Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In E. Harmon-Jones (Ed.), *Cognitive dissonance: Reexamining a pivotal theory in psychology (2nd ed.)* (pp. 3–24). American Psychological Association.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. *Methodology in the social sciences*. The Guilford Press.

Hertz, N., & Wiese, E. (2019). Good advice is beyond all price, but what if it comes from a machine? *Journal of Experimental Psychology: Applied*, *25*(3), 386–395.

Hollen, P. J. (1994). Psychometric properties of two instruments to measure quality decision making. *Research in Nursing & Health*, *17*(2), 137–148.

Homburg, C., Wieseke, J., & Hoyer, W. D. (2009). Social Identity and the Service-Profit Chain. *Journal of Marketing*, *73*(2), 38–54.

Jörling, M., Böhm, R., & Paluch, S. (2019). Service Robots: Drivers of Perceived Responsibility for Service Outcomes. *Journal of Service Research*, *22*(4), 404–420.

Kirby, R., Forlizzi, J., & Simmons, R. (2010). Affective social robots. *Robotics and Autonomous Systems*, *58*(3), 322–332.

Kyndt, E., Dochy, F., Struyven, K., & Cascallar, E. (2011). The perception of workload and task complexity and its influence on students' approaches to learning: a study in higher education. *European Journal of Psychology of Education*, *26*(3), 393–415.

Lankton, N., McKnight, D. H., & Tripp, J. (2015). Technology, Humanness, and Trust: Rethinking Trust in Technology. *Journal of the Association for Information Systems*, *16*(10), 880–918.

Lee, J., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Lüdtke, A., & Möbus, C. (2005). A Case Study for Using a Cognitive Model of Learned Carelessness in Cognitive Engineering. In G. Salvendy (Ed.), *Hci International 2005: 11th International Conference on Human-Computer Interaction*. Erlbaum 2005.

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association: JAMIA*, *24*(2), 423–431.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, *7*(3), 297–337.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *11th Australasian Conference on Information Systems*, *53*.

Manzey, D., Bahner, J. E., & Hueper, A.-D. (2006). Misuse of Automated Aids in Process Control: Complacency, Automation Bias and Possible Training Interventions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(3), 220–224.

McKibbon, K. A., & Fridsma, D. B. (2006). Effectiveness of clinician-selected electronic information resources for answering primary care physicians' information needs. *Journal of the American Medical Informatics Association: JAMIA*, *13*(6), 653–659.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research*, *13*(3), 334–359.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, *6*(1), 44–58.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In *Parasuraman, Mouloua (Eds.) 1996-Human factors in transportation* (pp. 201–220).

Mosier, K. L., Skitka, L. J., Burdick, M. D., & Heers, S. (1996). Automation Bias, Accountability, and Verification Behaviors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *40*(4), 204–208.

Mosier, K. L., Skitka, L. J., Dunbar, M., & McDonnell, L. (2001). Aircrews and Automation Bias: The Advantages of Teamwork? *The International Journal of Aviation Psychology*, *11*(1), 1–14.

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance Consequences of Automation-Induced 'Complacency'. *The International Journal of Aviation Psychology*, *3*(1), 1–23.

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39*(2), 230–253.

Parikh, N. (2021). Understanding Bias In AI-Enabled Hiring. *Forbes,* 2021. https://www.forbes.com/sites/forbeshumanresourcescouncil/2021/10/14/understanding-bias-in-ai-enabled-hiring/?sh=5fbf64427b96

Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, *66*, 29–38.

Raven, B. H. (2008). The Bases of Power and the Power/Interaction Model of Interpersonal Influence. *Analyses of Social Issues and Public Policy*, *8*(1), 1–22.

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. In C. Bartneck (Ed.), *The Eleventh ACMIEEE International Conference on Human Robot Interaction* (pp. 101–108). IEEE Press.

Shah, S. J., & Bliss, J. P. (2017). Does Accountability and an Automation Decision Aid's Reliability Affect Human Performance in a Visual Search Task? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 183–187.

Simon, M., Houghton, S. M., & Aquino, K. (2000). Cognitive biases, risk perception, and venture formation. *Journal of Business Venturing*, *15*(2), 113–134.

Skitka, L. J., Mosier, K. L., & Burdick, M. D. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991–1006.

Skitka, L. J., Mosier, K. L., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, *52*(4), 701–717.

Smids, J., Nyholm, S., & Berkers, H. (2020). Robots in the Workplace: a Threat to—or Opportunity for—Meaningful Work? *Philosophy & Technology*, *33*(3), 503–522.

Stock, R. M., Jong, A. de, & Zacharias, N. A. (2017). Frontline Employees' Innovative Service Behavior as Key to Customer Loyalty: Insights into FLEs' Resource Gain Spiral. *Journal of Product Innovation Management*, *34*(2), 223–245.

Vanden Abeele, M. M. P., & Postma-Nilsenova, M. (2018). More Than Just Gaze: An Experimental Vignette Study Examining How Phone-Gazing and Newspaper-Gazing and Phubbing-While-Speaking and Phubbing-While-Listening Compare in Their Effect on Affiliation. *Communication Research Reports*, *35*(4), 303–313.

Whyte, G. (1991). Diffusion of responsibility: Effects on the escalation tendency. *Journal of Applied Psychology*, *76*(3), 408–415.

Wiener, E. L. (1981). Complacency: Is the term useful for air safety. *Proceedings of the 26th Corporate Aviation Safety Seminar*, *117*, 116–125.

Wilson, P., Henry, J., Andersson, G., Hallam, R., & Lindberg, P. (1998). A critical analysis of directive counselling as a component of tinnitus retraining therapy. *British Journal of Audiology*, *32*(5), 273–286.

Yam, K. C., Bigman, Y., Tang, P. M., Ilies, R., Cremer, D. de, Soh, H., & Gray, K. (2020). Robots at work: People prefer-and forgive-service robots with perceived feelings. *The Journal of Applied Psychology*, *106*(10), 1557–1572.

Yeh, M., & Wickens, C. D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, *43*(3), 355–365.

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research*, *37*(2), 197–206.