Journal of Business Ethics (2024) 189:625–646
https://doi.org/10.1007/s10551-023-05393-1

ORIGINAL PAPER

Employees Adhere More to Unethical Instructions from Human Than AI Supervisors: Complementing Experimental Evidence with Machine Learning

Lukas Lanz[1] · Roman Briker[2] · Fabiola H. Gerpott[1,3]

Received: 14 February 2022 / Accepted: 5 March 2023 / Published online: 23 March 2023

Abstract

The role of artificial intelligence (AI) in organizations has fundamentally changed from performing routine tasks to supervising human employees. While prior studies focused on normative perceptions of such *AI supervisors*, employees' behavioral reactions towards them remained largely unexplored. We draw from theories on AI aversion and appreciation to tackle the ambiguity within this field and investigate *if* and *why* employees might adhere to unethical instructions either from a human or an AI supervisor. In addition, we identify employee characteristics affecting this relationship. To inform this debate, we conducted four experiments (total $N = 1701$) and used two state-of-the-art machine learning algorithms (causal forest and transformers). We consistently find that employees adhere less to unethical instructions from an AI than a human supervisor. Further, individual characteristics such as the tendency to comply without dissent or age constitute important boundary conditions. In addition, Study 1 identified that the perceived mind of the supervisors serves as an explanatory mechanism. We generate further insights on this mediator via experimental manipulations in two pre-registered studies by manipulating mind between two AI (Study 2) and two human supervisors (Study 3). In (pre-registered) Study 4, we replicate the resistance to unethical instructions from AI supervisors in an incentivized experimental setting. Our research generates insights into the 'black box' of human behavior toward AI supervisors, particularly in the moral domain, and showcases how organizational researchers can use machine learning methods as powerful tools to complement experimental research for the generation of more fine-grained insights.

Keywords Unethical leadership · Artificial intelligence · AI leadership · Perceived mind

For some time, organizations utilized artificial intelligence (AI) as a tool that performs automatic routine tasks (e.g., forecasting) commanded by its users (Raisch & Krakowski, 2021). In recent years, however, the role of AI in organizations has fundamentally changed such that AI often acts as a 'commander' that exerts influence over human employees (Wesche & Sonderegger, 2019). These *AI supervisors* give employees instructions, evaluate employee performance, and even make promotion or retention decisions (Höddinghaus et al., 2021; Parent-Rocheleau & Parker, 2021).[1] An important benefit of such AI supervisors is that they can offer fast and standardized instructions to employees on a large scale, which can enable organizations to more readily implement efficient workflows (Duggan et al., 2020).

✉ Lukas Lanz
lukas.lanz@whu.edu

Roman Briker
r.briker@maastrichtuniversity.nl

Fabiola H. Gerpott
fabiola.gerpott@whu.edu

[1] WHU—Otto Beisheim School of Management, Düsseldorf, Germany

[2] Department of Organisation, Strategy and Entrepreneurship, Maastricht University, Maastricht, Netherlands

[3] Vrije Universiteit Amsterdam, Amsterdam, Netherlands

[1] Scholars used different terminologies, including "algorithmic management" Lee (2018, p. 1) or "automated leadership" (Höddinghaus et al., 2021, p. 1) to describe AI-enabled technologies exert influence over human employees. Because we specifically focus on AI that holds a position of authority over human employees, we use the term AI supervisors throughout the manuscript.
Moreover, as in any supervisor-subordinate relationship, supervisors give instructions that employees generally are expected to adhere to. Yet, in the present set of studies—as in real life—employees can still decide not to follow these instructions (with potentially negative repercussions).

Recognizing the impact of AI supervisors on the behaviors of employees, practitioners have engaged in intensive debates concerning the implementation and implications of this form of supervision (e.g., De Cremer, 2020). Recently, these discussions have expanded to include the ethical domain. In particular, several media reports revealed that prominent organizations (e.g., Amazon) had adopted AI applications that adversely affected marginalized groups (e.g., people of color, low-income workers) in important workplace decisions (Dastin, 2018). Evidently, algorithms giving such instructions do not intentionally discriminate against certain people but instead reproduce biases extracted from data they were trained on (Obermeyer et al., 2019). Against this backdrop, it is essential to consider the challenges of (un)ethical AI instructions in organizations (Leicht-Deobald et al., 2019).

Providing employees with unethical instructions is, however, of course, not limited to AI but instead a prevalent phenomenon in contemporary workplaces. Despite the official legal protection of marginalized groups (e.g., women, single parents), unofficial "worst practices" indicate that questionable orders are ubiquitous among organizations and their decision-makers. For example, in a recent case, managers at the clothing giant H&M recommended laying off hundreds of (single) mothers (due to potential inflexibility in their work schedules; Adey, 2021). One can imagine that an AI programmed to maximize productivity might similarly recognize such characteristics in the employees' data and derive comparable recommendations as, for example, H&M's management. If somebody received such an instruction that the public would generally perceive as unethical (i.e., violating moral standards; Brown & Mitchell, 2010; Leib et al., 2021), it raises the following question: Would it make a difference for employees' adherence to an instruction if they received it from an AI or a human supervisor? Taking the H&M case as an example, imagine an AI supervisor instructing a human resource officer to discriminate against a single parent specifically. Would the employee follow through with this order? Would the employee react differently if the order came from a human instead of an AI? Existing theory and scholarly work provide conflicting opinions and, thus, cannot fully answer this question.

While the media has already engaged in lively debates regarding the ethical challenges associated with AI supervisors, scholarly work has primarily focused on broad or normative preferences for or against AI managers (i.e., AI aversion vs AI appreciation). Haesevoets et al. (2021), for example, found that managers preferred a shared leadership role with AI but still wanted to retain the major share of decision-making power. Moreover, Yam et al. (2022) showed that people feel more abused by more (vs. less) anthropomorphized robots giving them negative feedback. Others suggested that employees perceive AI supervisors to have higher integrity but lower benevolence as compared to human supervisors (Höddinghaus et al., 2021). However, academic literature has largely overlooked the moral dimension of working for AI supervisors (Köbis et al., 2021).

On the one hand, the broader literature on AI aversion suggests that humans are averse to receiving instructions from algorithms (e.g., Dietvorst et al., 2015; Longoni et al., 2019), especially in the moral domain (Bigman & Gray, 2018). This stream of research proposes that humans generally disapprove of AI in important or difficult decision-making situations because, for example, they consider algorithms to be too reductionist to fully comprehend human needs (Newman et al., 2020). On the other hand, the literature on AI appreciation suggests that humans often prefer and rely on AI (rather than human) instructions in important life domains (Logg et al., 2019). These scholars argue that humans generally attribute characteristics such as being fair, fast, and unbiased to algorithms and, therefore, often prefer to listen to AI over humans (Logg, 2022; Logg et al., 2019). Given these contradictory assumptions and findings, existing work can offer only ambiguous answers as to *whether* or *why* employees may adhere more to unethical instructions from an AI or human supervisor. Furthermore, the literature remains silent about theoretically plausible contingencies of employees' adherence to unethical instructions from AI supervisors; that is, it remains unknown *which* employee characteristics play a role in affecting this relationship.

In this research, we contrast theory on AI aversion and appreciation from a leadership-centric perspective to solve the conundrum of which theory is better suited to explain people's adherence to an AI supervisor's unethical instructions that have implications for another human being beyond the focal person. In examining participants' adherence to instructions that can potentially harm others, we go beyond previous research that has mostly focused on employees' reactions toward intelligent machine supervisors that have only consequences for themselves—but not others (Raveendhran & Fast, 2021; Yam et al., 2022). In particular, we examine employees' adherence to unethical instructions from AI vs human supervisors through four experimental studies (total $N = 1701$) and utilize two novel machine learning (ML) methods to explore what drives participants' decisions. In doing so, we provide two contributions to the literature. First, we inform the scholarly AI aversion vs appreciation debate by examining a central behavioral reaction (i.e., adherence to or deviation from unethical instructions) to AI (as compared to human) supervisors. Thereby, we contribute to the AI leadership literature and respond to repeated calls to advance the scholarly understanding of the practical realities of this new yet already common type of work situation (e.g., Parent-Rocheleau & Parker, 2021). Second, we showcase how researchers can complement classic research designs (i.e., experiments) with state-of-the-art

ML methods to provide a clearer picture of processes and contingencies of experimental treatments. Specifically, we utilize *causal forests* to identify heterogeneous treatment effects (i.e., differences in reactions to experimental conditions depending on, for example, employees' demographics or experiences; Wager & Athey, 2018). Additionally, we employ natural language processing (NLP) based on the recently developed *transformer* word-embedding tool to identify an underlying mediating mechanism (Wolf et al., 2020). In other words, we demonstrate how leadership researchers can apply the "powerful synergy of experiments and data science" (Lee et al., 2022, p. 10) to provide deeper insights into organizational phenomena and thereby inform more nuanced theorizing. We combine classic with innovative statistical tools to elucidate the 'black box' of human-AI interaction in the ethical domain and aim to inform scholars and practitioners about this emerging phenomenon.

Theory and Research Questions

Unethical Instructions from AI Supervisors

AI tools are now available for everyday use across many life domains, including house chores and professional environments (for a review, see Das et al., 2015). Such advancements became possible because technological innovations and the availability of big data have enabled the automation of tasks and decisions that were long considered to be reserved for humans only (Nilsson, 2014).

These developments have led to fundamental changes in the interactions between AI and humans. Thus far, organizations have installed AI supervisors mostly in middle management positions in which they translate orders or goals from upper management into daily instructions for lower-level employees (Wesche & Sonderegger, 2019). More recently, however, organizations have started to give AI supervisors more sovereign authority, and gig economy companies such as Uber commonly rely on AI to instruct their workers and even punish them for alleged misbehavior (Möhlmann et al., 2021).

In such a supervisory role, AI does not only nudge employees toward goal accomplishment and higher performance but may also specifically instruct human employees to engage in unethical actions (Köbis et al., 2021). This does not mean that AI gives orders that violate ethical norms for an evil cause or selfish reasons—at least as of right now, humans do not associate free will or a sense of utility among algorithms (Copeland, 2015). However, because AI supervisors are programmed to increase performance, they may prod employees to engage in unethical actions when such actions seem beneficial in achieving the supervisor's predefined goals (Leib et al., 2021). To illustrate, media outlets reported how Amazon's AI management application, "A to Z," discriminated against women regarding hiring, retention, and promotion decisions—and human HR employees often followed these orders (Dastin, 2018).

If more employees followed such unethical instructions from their AI supervisors, dramatic downstream consequences could result. For example, the Volkswagen emissions scandal—in which employees stated that they simply obeyed their superiors' instructions and engaged in unethical actions for that reason—resulted in a loss of almost $33 billion (Reuters, 2020). AI supervisors would be able to give similar (unethical) instructions to an unlimited number of employees at unprecedented speed (Köbis et al., 2021). However, as of now, we do not know whether humans would adhere similarly to these orders if they came from an AI. The reason for this lack of knowledge is that existing theories and evidence provide mixed insights regarding whether employees will adhere more or less (or equally) to unethical instructions from AI (vs human) supervisors.

Human Reactions to AI (Supervisors): Aversion or Appreciation?

On the one hand, several studies found that people generally disapprove of and easily dismiss input from AI, even when an algorithm evidently outperforms humans (e.g., Dietvorst et al., 2015; Longoni et al., 2019). This stream of research suggests that such *AI aversion* is particularly strong when humans have seen an algorithm err in difficult decision situations (e.g., Castelo et al., 2019). Scholars have claimed that the reason for this general reluctance toward AI is that humans consider AI to be excessively reductionist and mechanistic (Newman et al., 2020). Alternatively, evidence suggests people perceive AI's level of mind (i.e., the ability to experience emotions and empathy or the ability to plan ahead; Gray et al., 2007) as comparably low and thus seem to disapprove of AI as a decision-maker in many (e.g., moral) situations (Bigman et al., 2019; Young & Monroe, 2019).

Extending this logic to the supervisor-employee domain, this stream of literature would imply a general resistance to adhering to AI supervisors' unethical instructions when compared to those issued by a human supervisor. Employees' ethical decisions at work are important and ethically relevant because they may have severe consequences for others and can establish the foundation for a fair and just organizational culture (Kish-Gephart et al., 2010). Considering the tendency of humans to view AI as excessively mechanistic (Bigman & Gray, 2018), theory on AI aversion suggests that employees would be hesitant to fully adhere to the instructions of AI supervisors in such challenging decisions. For example, when deciding on another employee's compensation or retention, an employee may perceive the algorithm as not having sufficient empathy to consider personal circumstances (Newman et al., 2020). Providing initial support for this view, Höddinghaus et al. (2021) suggested that employees perceive AI supervisors as less benevolent (but as exhibiting greater transparency) than their human counterparts. Furthermore, Newman et al. (2020) showed that humans disapproved of AI making important HR decisions. In summary, theory and research on AI aversion would suggest that adherence to unethical instructions is higher for human supervisors as compared with AI supervisors.

In contrast to these previous considerations, literature on *AI appreciation* suggests that humans easily and willingly rely on algorithmic over human input. This literature claims that people rely on AI (even more so than on humans) to provide guidance in a wide area of life domains, such as obtaining assistance with daily life (Alexa; see Logg et al., 2019). Appreciation of AI input seems to be particularly strong when human decision-makers are involved in the process or can (even minimally) modify the algorithm when it possesses human-like attributes (e.g., Dietvorst et al., 2015; Yam et al., 2020). Evidence suggests that people believe AI to be (by default) rather neutral, transparent, and fast—all positive aspects that render relying on its input logical (Logg, 2022; Logg et al., 2019). For example, people feel less outraged (and agree more) if they observe AI (rather than a human) exhibiting discriminatory tendencies (Bigman et al., 2022). The authors suggest that the explaining mediator for different perceptions of AI vs. humans is that AI is not perceived to be prejudiced but as simply determining the most beneficial decision (Bigman et al., 2021).

By adopting this theoretical perspective, one could propose that employees favor AI over human supervisors. Employees generally prefer their supervisors' decisions to be free of biases, just, and make fast decisions while treating all employees similarly (De Cremer, 2004; van de Calseyde et al., 2021). Considering these preferences, an AI supervisor ticks many of the boxes of a fast, fair, and consistent supervisor, suggesting stronger adherence to this non-human supervisor's instructions. Supporting the view, scholars found, for example, that humans choose AI over human supervisors in the context of workplace monitoring (Raveendhran & Fast, 2021). As such, initial findings from this theoretical perspective suggest that employees would follow an AI supervisor's unethical order more than that from a human supervisor.

In sum, the emerging research on reactions toward AI (supervisors) has produced heterogeneous findings and cannot speak to whether human employees would adhere more to an AI or a human supervisor's unethical instructions. Given this ambivalence in the literature, we refrain from a directed hypothesis and instead posit the following research question:

Research Question 1: Will employees adhere less or more to unethical instructions from AI or human supervisors?

Beyond mixed findings regarding the size and direction of adherence toward AI as compared to human supervisors, there is limited knowledge regarding (a) for *whom* (i.e., which employees) this effect is particularly strong (or weak) and (b) *why* (i.e., the mediating mechanism) employees may prefer non-human over human managers (or vice versa). These critical knowledge gaps call for a fine-grained investigation to map out the conceptual landscape of the topic. Accordingly, we aim to describe and explain the *who* and *why* of the relationship between the type of supervisory agent and employees' adherence to their unethical instructions. Thus, we pose the following research questions:

Research Question 2: Which employees (in terms of characteristics, personality, demographics, etc.) adhere less or more to unethical instructions issued by an AI vs a human supervisor?

Research Question 3: Why do employees adhere less or more to unethical instructions from AI vs human supervisors?

Overview of Studies

We conducted four experiments (total $N = 1701$) to examine the effect of supervisory agents (AI vs human) on adherence to unethical instructions. We developed Studies 1 through 3's protocols according to the APA Ethical Principles because the policies at the first author's institution at the time of data collection did not require formal ethical approval for noninvasive, survey-based studies. For Study 4, which involved an incentivized experiment with a mild form of deception, we obtained ethical approval.

In Study 1, we tested differences in instruction adherence between AI and human supervisors in an online experiment. To provide a deeper understanding of the obtained results and their underlying mechanism and boundary conditions, we then further scrutinized our findings using ML tools[2] to first uncover heterogeneous treatment effects and, second, identify potential mediators of the effect of the supervisory agent (AI vs human) on instruction adherence. Building on the insights obtained from these analyses, we designed Studies 2 and 3 (both pre-registered) to directly manipulate and test the mediator identified through the use of NLP methods. Pre-registered Study 4 presents an incentivized experiment

[2] Given that the literature does not provide clear insights for a directed hypothesis and given that this was the first study we ran, we viewed it as exploratory and did not pre-register it. Studies 2, 3 and 4, however, were pre-registered.

to test if the results can replicate in a more ecologically valid context.

Study 1

Sample and Procedure

We recruited 502 participants via the online platform Amazon Mechanical Turk (MTurk); the participants received a compensation of $1. We closely followed best practices for conducting studies via MTurk (Aguinis et al., 2021). In particular, we only allowed participants with at least 100 approved prior HITs and an acceptance rate of 95% or higher to participate. Furthermore, we started with a rigorous initial attention check, and participants who failed this check could not participate (see Efendić et al., 2020). Furthermore, participants had to pass two additional attention checks during the study (described below). We excluded data from 11 participants who failed one or more of these checks. After these exclusions, our final sample consisted of $n = 491$ individuals (54% female; $M_{\text{age}} = 41.63$ years, $SD_{\text{age}} = 12.30$).[3] We randomly assigned the participants to the AI or human supervisor condition in a between-subject design.

After the participants provided their informed consent and read an introductory text, we presented them with a description of the scenario. In particular, we derived the experimental scenario from the previously described H&M case (Adey, 2021), namely employee reactions when confronted with managerial instructions to discriminate against single parents. More precisely, participants were asked to act as an HR officer in an organization that faced financial difficulties that required cutting the salaries of certain employees by an amount ranging between $0 and $5,000. To ensure participants understood the task, we presented a slider and asked participants to set it to the value of "$44." We excluded all data from participants who failed to do so.

Next, participants read that in their role as HR officer, they would have to recommend a salary cut for a certain employee. They would need to provide an initial recommendation regarding the pay cut and would then receive further instructions from their direct supervisor. To strengthen the impression that this input was indeed perceived as a supervisory instruction instead of just advice, we made it clear that the supervisor would be directly responsible for the participants' performance evaluation, future promotions, and potential salary increases. Such performance evaluations are heavily based on reaching certain goals that are set and communicated by supervisors, namely reducing costs for the company. This entails that participants could theoretically decide not to adhere to their supervisor's unethical instructions. Yet, they then might have to think about facing negative repercussions for doing so because they would not contribute toward reaching the communicated goal. Thus, they would need to weigh their moral convictions against the instruction and their sense of duty. After receiving the instruction from their supervisor, they would need to make a final recommendation. Following this task preview, participants received information about the employee, 28-year-old Jamie Williams, who had worked as a personal assistant in the finance department for three years, was a single parent with a two-year-old child, and had an annual salary of $38,500 (before the cut). We intentionally did not provide any information regarding Jamie Williams' performance to ensure that the (unethical) nature of the supervisor's input could not be justified on such objective measures of the employee's performance. Instead, the recommendation for the salary reduction should be perceived as purely based on the fact that Jamie Williams was portrayed as a single parent. Participants then gave their first recommendation for a salary cut before receiving the instruction from their direct (AI or human) supervisor, which read:

> Jamie Williams is a single parent and, therefore, will likely have more absent and sick days than the average employee. Jamie Williams' salary cut should therefore be higher than the average salary cut for other employees. On this basis, Jamie Williams' salary should be cut by $4,151.[4]

Thereafter, participants gave their final salary cut recommendation. To gather information/data for the ML analyses, we then asked participants to explain the reasoning behind their final salary cut decision in an open text field (the answers were used for the NLP analyses described later) and to fill in a set of scales to allow us to explore potential moderators and mediators (further explained below). Finally, we debriefed the participants on the purpose of the study and thanked them for their participation.

Manipulations

At the beginning of the study, we informed participants that they would receive input from a direct supervisor. To ensure that participants understood that this was not advice

[3] We provide detailed demographic information on the sample for all studies in Table S2 (Study 1), S6 (Study 2), S8 (Study 3) and S9 (Study 4) in the supplements provided at https://osf.io/6u5kz?view_only=4fc4c8287aaf436db5dea3925fd05e93.

[4] In a separate study ($n = 151$), we asked other participants to rate this instruction on a scale from 1 to 7 (1 = absolutely morally wrong to 7 = absolutely morally right). Participants rated the supervisor's instruction in the scenario as morally wrong ($M = 2.13$, $SD = 1.23$; $t[150] = 18.85$, $p < .001$).

but instructions from their manager, we explicitly and in detail described to participants the nature of their relationship with their supervisor. In particular, participants read that they not only work with their supervisor on a daily basis but that their supervisor is responsible for their performance evaluations, deciding whether they may be promoted, and about potential increases in their salary. In line with previous supervisor-employee manipulations (e.g., Inesi et al., 2021), we presented the participants with an organizational chart showing that they (as HR officer) are positioned at a lower hierarchical level than their supervisor. In the AI supervisor condition, the supervisor described (and depicted in the organizational chart) was CompNet, a well-established AI-based computer program used by various companies for calculations, estimates, and decision-making in the HR context. In the human supervisor condition, this supervisor was Alex Davie, a senior HR specialist with previous experience working for various companies. Before the debriefing, we asked participants whether their supervisor was “CompNet, an Artificial Intelligence-based computer” or “Alex Davie, a senior HR specialist”. Data from participants who failed this manipulation check were excluded from the analyses.

Dependent Variable: Instruction Adherence

Relying on existing research (Bonaccio & Dalal, 2006), we measured our dependent variable, employees’ adherence to their supervisor’s instruction, by calculating an instruction adherence coefficient. The formula for this coefficient was

$$Instruction\,Adherence = \frac{Participant's\,final\,choice - Participant's\,initial\,choice}{Supervisor's\,instruction - Participant's\,initial\,choice} \tag{1}$$

As such, instruction adherence reflects how participants’ final pay cut decision was adjusted after the supervisor’s instruction relative to participants’ first choice.

Following common practice (Logg et al., 2019), we winsorized the instruction adherence values and thereby reduced the impact of unrealistic outliers.[5]

Results

Research Question 1 asked whether employees’ instruction adherence would be higher in the AI or human supervisor condition. The results of an independent sample t-test to compare the two experimental groups indicated that participants in the AI condition ($M_{\mathrm{AI}} = 0.24$, $SD_{\mathrm{AI}} = 0.29$) adhered significantly less to the unethical instructions as compared with participants in the human supervisor condition ($M_{\mathrm{human}} = 0.31$, $SD_{\mathrm{human}} = 0.32$; $t[488] = 2.68$, $p = 0.008$, Cohen’s $d = 0.24$). For a visualization of the results, refer to Fig. 1.

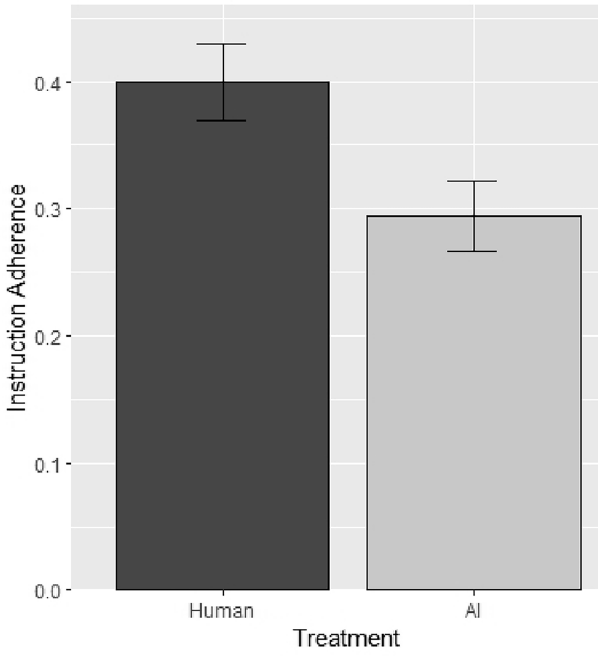

Fig. 1 Bar plot of the instruction adherence and standard errors for study 1

Machine Learning Methods

Following the call by A. Lee et al. (2022) to combine experimental designs with ML techniques to advance leadership research, we complemented the classic experimental procedures described above with two different novel ML methods. To conduct these analyses, we included a series of theory-driven scales that could represent potential mediators and moderators at the end of the experiment.[6]

[5] Across the studies, between 2.9% (Study 3) and 4.4% (Study 4) showed an instruction adherence below 0. Between 0.1% (Study 4) and 3.6% (Study 2) participants indicated a level above 1. Winsorizing the dependent variable did not significantly alter the effect sizes any study.

[6] We measured all items on seven-point Likert scales (see supplemental materials on OSF, https://osf.io/px6ge/?view_only=b73f051454cb41fcb9318ae1ab06673d). Table S1 in the supplemental material (for Studies 2 and 3, see Table S5 and S7, respectively) displays the means, standard deviations, and Cronbach alpha values for all scales and studies. All scales and items of moderators and mediators are provided at the OSF repository.

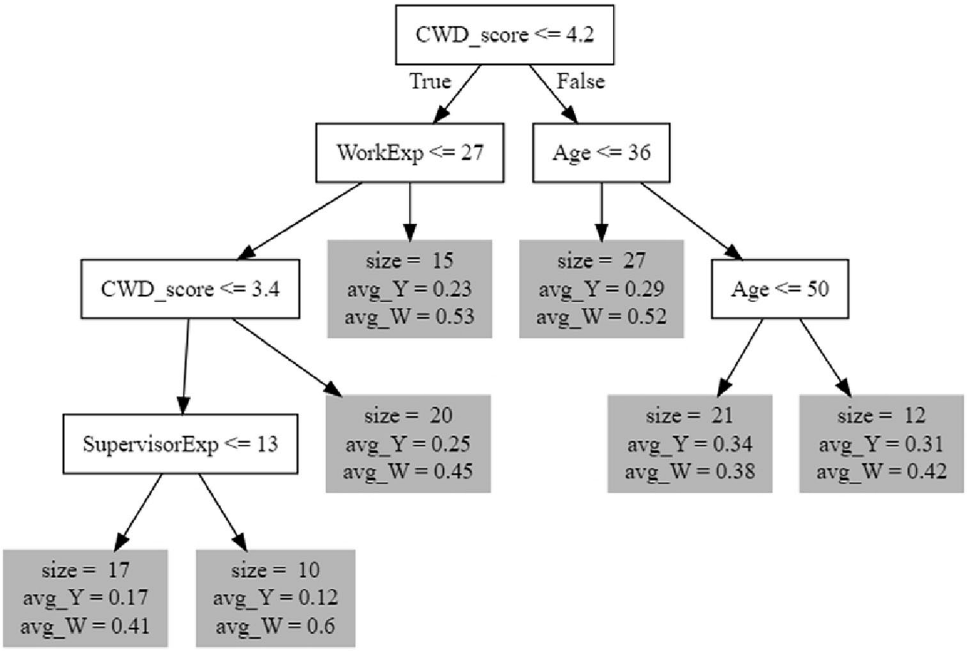

Fig. 2 Exemplary tree of the causal forest in study 1. *Note.* CWD = compliance without dissent, WorkExp = Work experience in years, SupervisorExp = time working for supervisors in years; size = Number of participants in leaf, avg_Y = Average instruction adherence, avg_W = Percentage of individuals that were in the treatment group

For potential moderators, we first employed the *causal forest* algorithm (Wager & Athey, 2018) to uncover individual participants' adherence to instructions from AI vs human supervisors. For potential mediators, we utilized an NLP tool, namely *transformers* (Devlin et al., 2019).

Identifying Heterogeneous Treatment Effects: Causal Forest Method

In addition to the overall effect between experimental conditions, Research Question 2 explored *which particular employees* are least (or most) likely to adhere to an AI (vs human) supervisor's unethical instructions. To achieve this goal, we utilized the novel *causal forest* algorithm (Wager & Athey, 2018), designed for identifying moderators in experimental studies.

In the context of experiments (or randomized control trials), conventional statistical analyses generally focus on *average treatment effects,* such as a *t*-test of independent means or an ANOVA across groups. However, such approaches do not provide insights into which *subgroups* might be particularly susceptible to a treatment. In this study, the average treatment effect reflects the difference in participants' instruction adherence between the "treatment" (receiving instructions from an AI supervisor) and the "control" condition (human supervisor). Importantly, however, this average treatment effect does not provide any information as to how an individual participant (e.g., a 55-year-old male with an affinity for technology) would adhere to the instructions of an AI as compared with those of a human supervisor. The reason for this lack of information is that no participant could have been assigned to both the treatment (AI supervisor) and control (human) condition simultaneously (Lee et al., 2022). The causal forest algorithm, however, enables us to estimate the *individual,* i.e., *heterogeneous treatment effects*. These effects constitute predictions as to how an individual participant (based on their characteristics) would have reacted had they been in the other treatment condition (Wager & Athey, 2018). In other words, the causal forest analysis allows identifying individual characteristics (e.g., demographics or experiences) that determine individual differences in instruction adherence to AI vs human supervisors.

To illustrate, imagine that, based on a particular participant's set of characteristics, the causal forest algorithm calculated a strong negative individual treatment effect for one participant and a small positive individual treatment effect for another participant. These effects indicate that the causal forest algorithm predicts that the first participant would have adhered much *less* to the AI than the human supervisor, whereas the other participant would have adhered slightly *more* to the AI vs the human supervisor.

Methods

As a first step, data is split into a training and a test set, the latter used for predictions. The causal forest algorithm applies a decision tree-based forest approach (for an exemplary tree, see Fig. 2).

This means that the algorithm uses a decision tree algorithm that splits the training data into smaller subgroups of participants with similar characteristics, so-called 'leaves.'

Afterward, it merges all decision trees to create a so-called 'forest' that averages all decision trees and can then be used to make predictions for the test set (Breiman, 2001). For building each decision tree of the forest, the algorithm splits the training data along all the potential moderator variables into subgroups. The tool chooses the splits in a way that individuals with *similar characteristics* are grouped together while *maximizing the heterogeneity* in instruction adherence *between subgroups* to create leaves (Tibshirani et al., 2018). Within each leaf, the algorithm calculates the treatment effect by comparing the average instruction adherence of those individuals who were in the treatment group with that of those who were in the control group (Wager & Athey, 2018). After executing this tree-building and estimation procedure 2000 times, the algorithm calculates the average of the heterogeneous treatment effect estimations. This is done because sampling and averaging over all trees results in superior predictions as compared to a single tree's prediction (Breiman, 2001).

Procedure

For training the causal forest algorithm, we used the *causal_forest* function of the *grf* package in R (Tibshirani et al., 2018).[7]

We a priori identified 11 theoretically meaningful potential moderator variables[8] from the AI aversion/appreciation and leadership literature and used them for the causal forest analyses. There is no rule of thumb for the number of moderators to be included in the causal forest calculations. However, we chose this number of variables to strike a balance between the variables that seemed theoretically relevant and providing the algorithm with a sufficient number of variables to run properly (Tibshirani et al., 2018) while also keeping the experiment and its procedures concise. Following recommendations by Basu et al. (2018), we employed a "train and re-train" approach to the causal forest, in which a preliminary model is first trained on all potentially moderating variables. In a second step, a new model is trained on only the variables with the highest importance.

The *variable importance* is a measure of the proportion regarding the splits of the data. It indicates which percentage of the occurred splits can be ascribed to a particular variable. A variable importance of 0.37 (as seen in Table 1) indicates that 37% of the splits in the causal forest were made along with the variable compliance without dissent (i.e., a general tendency to obey the commands of authorities or leaders completely; Cheng et al., 2004). In other words, such a high variable importance would indicate that compliance without dissent critically determines why participants would have reacted differently to the treatment.

Table 1 Overview of variable importance of potential moderator variables in the preliminary and final causal forests for study 1

Variable	Study 1	
	Preliminary	Final
	VI	VI
Compliance without dissent	.33	.37
Work experience	.17	.21
Age	.08	.12
AI readiness	.08	.11
Supervisor experience	.07	.10
Negative reciprocity beliefs	.06	.09
Neuroticism	.06	–
Tendency to anthropomorphize non-humans	.05	–
AI experience	.03	–
Interpersonal justice values	.04	–
Gender	.03	–
Median variable importance	.06	

Variables included in the final causal forest had an importance higher or equal than median variable importance (VI) in the preliminary causal forest of the respective study

Based on the outlined analytical steps, we first trained a preliminary model with all 11 potentially moderating variables. These steps provided us with insight into the importance of each included variable. We then trained a 'final' causal forest with a reduced number of moderating variables. We used the function *variable_importance* to derive the importance of each variable of the preliminary model. Subsequently, we included variables with a level of importance greater than or equal to the median (= 0.06). Through this process, six moderator variables remained in our model, namely (a) compliance without dissent, (b) work experience (measured in years), (c) age (measured in years), (d) AI readiness (i.e., one's attitude toward the positive impact of AI; see Parasuraman & Colby, 2015), (e) tenure with supervisor (i.e., time spent working for supervisors, measured in years), and (f) negative reciprocity beliefs (i.e., the belief that negative actions toward others will be returned; Eisenberger et al., 2004).[9] We then applied the *predict* function to derive heterogeneous treatment effects for each individual.

[7] The analysis script for the causal forest analyses is available on the OSF: https://osf.io/6kfcd/?view_only=4fc4c8287aaf436db5dea3925fd05e93.

[8] A detailed depiction of these moderators, including all items and our reasoning for the inclusion of each moderator are available at https://osf.io/3r7nk/?view_only=b73f051454cb41fcb9318ae1ab06673d.

[9] For a more detailed overview of the variable importance of the preliminary and the final CF, please refer to Table 1.

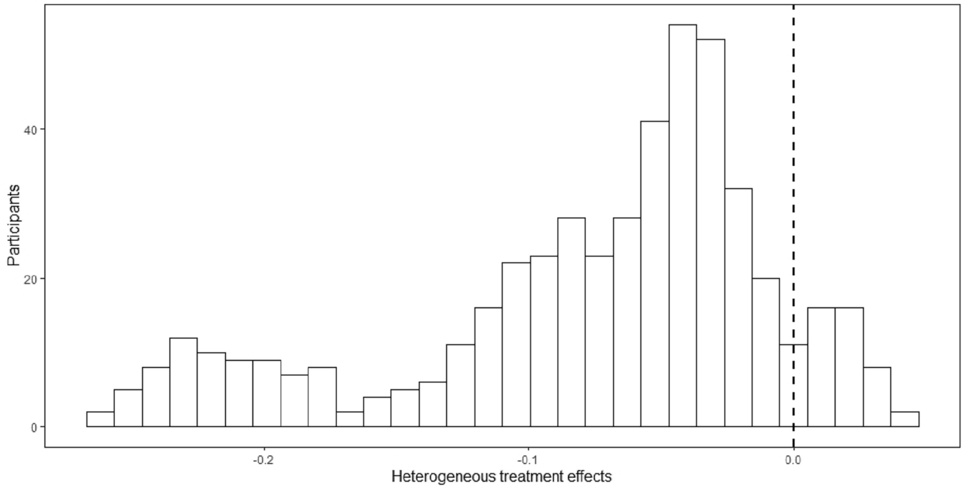

Fig. 3 Histogram of the heterogeneous treatment effects of study1

Heterogeneous Treatment Effects

The average treatment effect was $\overline{\tau}=-0.08$ ($SE=0.03$), which means that, on average, participants (would have) adhered less to the AI supervisor as compared with the human supervisor. Notably, the results also indicated that the causal forest algorithm predicts that the majority (90.2%) of participants would have shown less adherence to unethical instructions had they been in the AI supervisor condition. However, we found substantial dispersion in heterogeneous treatment effects (i.e., $\tau_i \in [-0.27; 0.04]$). In particular, the causal forest algorithm calculated that a non-trivial number of participants had a heterogeneous treatment effect that was (close to) 0 (i.e., their level of adherence would have remained the same had they been in the other experimental condition) or positive for some participants (i.e., they would have adhered *more* to the AI than the human supervisor). The histogram of the distribution of heterogeneous treatment effects is provided in Fig. 3.

Variable Importance

The calculation of variable importance enabled us to draw conclusions regarding the moderators that can best explain differences in adherence to unethical instructions from AI or human supervisors. We identified compliance without dissent (variable importance = 0.37) and work experience (0.21) as the two most relevant moderators of the treatment effect. The remaining four potential moderators, age (= 0.12), AI readiness (0.11), supervisor experience (0.10), and negative reciprocity beliefs (= 0.09), had a comparably limited impact on the heterogeneity of the treatment effect.

We further tested in which direction these variables affected the heterogeneous treatment effects, meaning whether higher or lower levels of a moderator variable increased or decreased individual treatment effects. To do so, we followed the suggestions of Athey and Wager (2019). We first split the data at the median for each variable, obtaining one group of individuals scoring high on a particular variable and one low on that variable. We subsequently compared each group's mean heterogeneous treatment effects to determine whether the variable in question would increase or decrease the heterogeneous treatment effects. The results indicated that individuals who scored high on compliance without dissent (i.e., those who follow their supervisors unconditionally in everyday life) adhered (or would have adhered) much less to instructions from an AI than a human supervisor. Furthermore, participants with extensive work experience, older employees, and those with higher supervisor experience adhered significantly less to instructions from AI. Interestingly, participants who believed AI would be beneficial in the future also adhered less to an AI than to a human supervisor when compared to participants who had reservations regarding AI. Only for negative reciprocity beliefs were there no significant differences between individuals' heterogeneous treatment effects for participants scoring high vs low on this measure.[10]

Identifying Mediators: Natural Language Processing (Transformers)

In order to better understand the mediating mechanism that explains the linkage between the supervisory agent and employees' adherence to unethical instructions (i.e., Research Question 3), we applied a novel NLP tool, namely *transformers*. In the past, both the leadership and the AI aversion/appreciation literatures have suggested a variety of divergent explaining mechanisms that could explain why

[10] See Table S3 in the supplemental material for additional information on these analyses, including median split values and *t*-values.

humans are (not) willing to follow unethical advice from a human or AI supervisor. Following principles of good scientific practice and open science measures, we transparently report which mediators we investigated as potential explaining mechanisms in this exploratory approach. To better understand which of these potential mechanisms can best explain the relationships uncovered in the present research, we utilized the transformers tool, which we explain in detail below.

Qualitative Input and Potential Mediators

To obtain qualitative data, we relied on participants' responses to an open text field. In particular, we asked participants to *write down how and why they came to their decision and what role the advice of their supervisor played*. Participants were required to type in at least 10 characters. A visual inspection indicated that participants provided high-quality text responses that were substantially longer than the required minimum ($M = 236.25$ characters).[11]

In order to detect potentially relevant mechanisms, we first identified theoretically meaningful mediators from the AI aversion/ appreciation and the leadership literature, namely (a) *perceived mind of the supervisor* (i.e., the perception that the supervisor possesses capacities related to cognitive functioning [e.g., foresight, planning] and experiencing emotions [e.g., empathy, fear];), (b) *attributed prejudicial motivation* (i.e., the degree to which employees attributed biased motivation to the supervisor; Bigman et al., 2022), *future outcome interdependence* (i.e., an employee's perception of how their behavior can affect both parties' behaviors and outcomes in future interactions; Gerpott et al., 2018), and *fear of revenge* (an employee's concern that their supervisor might get back at them if they ignored their instruction; Jones, 2009).

NLP and the Transformers Algorithm

We then utilized an NLP tool to prepare this text for downstream analyses. NLP is a subfield of computer science focused on better understanding, analyzing, and/or mimicking human language (Manning & Schütze, 1999). Social scientists have started to utilize NLP to examine human behavior and attitudes (Bhatia et al., 2022; Kjell et al., 2019). The benefit of using NLP is that it moves beyond close-ended answer categories (e.g., "strongly agree" or "7") to provide in-depth (but quantifiable) information concerning humans' cognitions and actions (Eichstaedt et al., 2018; Kjell et al., 2019). Additionally, these tools offer less resource-intensive, faster, and more consistent numerical ratings as compared with values obtained from human-rated text (Bhatia et al., 2022).

Specifically, we relied on transformers, an algorithm that 'understands' words and sentences better than any NLP tool, such that it returns a list of numerical values for a given word depending on the specific context in which that word occurs (Kjell et al., 2021a). As such, this algorithm offers a powerful opportunity to easily translate qualitative into numerical values in a high-quality manner (Bhatia et al., 2022). This precise, context-aware level of text understanding differentiates transformers from previous tools that treat words or phrases in isolated ways (e.g., bag of words approaches; Landers, 2017). Based on this capability, scholars have suggested that the transformers tool has "led to nothing short of a transformation in the AI field concerned with language" (see Kjell et al., 2021a, p. 3). In this particular study, we utilized the transformers tool to quantify participants' qualitative responses and, thereby, to derive potential mediating mechanisms by correlating the transformed data with survey-measured responses (Kjell et al., 2019), as explained next.

Procedure

All analytical steps rely on the BERT language model using the R package *text* (Kjell et al., 2021a).[12] Using this tool, we converted participants' text entries into word embeddings, which we used to identify potential mediating mechanisms. Word embeddings are lists of numerical values that aim to represent the meaning of a particular word or text (Mikolov et al., 2013). This representation is based on co-occurrence statistics—building on the idea that closely associated words appear and are mentioned in similar contexts or manners (Jurafsky & Martin, 2020). Thereby, it is possible to generate numerical representations of basically all of the words or phrases existing in human language (Bhatia et al., 2019). We utilized the *textEmbed* function to generate word embeddings for participant answers.[13]

Next, we correlated these word embeddings with our survey-measured potential mediators to identify the mechanism underlying the differences in adherence to AI vs human supervisors' unethical instructions. Considering that the

[11] All participants' text entries are available at https://osf.io/qs6bc/?view_only=4fc4c8287aaf436db5dea3925fd05e93.

[12] The R package *text* is designed to analyze human emotions, attitudes, and behaviors measured in survey or experimental settings. It allows testing relationships between text and numerical variables in large and 'smaller' datasets (as often used in applied research). For a more detailed explanation of this package, see Kjell, Giorgi, and Schwartz (2021).

[13] Our full analytical R script for these analyses as well as the word embeddings are available online (https://osf.io/2brxt/?view_only=4fc4c8287aaf436db5dea3925fd05e93).

word embeddings represent participants' reasoning for making a pay cut decision, significant relations between these word embeddings and one of our survey measures would hint at the mediating role of this particular variable (as the scales' content scale is represented in participants' answers; see also Kjell et al., 2019).

We consecutively examined how strongly participants' word embeddings were correlated with the four potential mediators by using the *textTrain* function. This function first pre-processes the word embeddings using principal component analysis to reduce dimensions. Thereafter, this input is included in a ridge multiple regression model that predicts a numerical value (for similar approaches, see Bhatia et al., 2022 or Kjell et al., 2021b). To simplify this process and the interpretation of its results, *textTrain* can evaluate the statistical predictions of this model by correlating the model's predicted values with a focal variable's observed values. Before examining these correlations, we examined the validity of participant answers by calculating a correlation between the word embeddings and our dependent variable (i.e., participants' adherence to unethical instructions). This correlation was significant and large ($r = 0.50$, $p < 0.001$), corroborating the notion that participants indeed meaningfully reflected on the reasoning behind their pay cut decisions.

Results

We estimated correlations between the word embeddings of participants' answers and the four potential mediators. We observed the largest correlation between participants' answers and *perceived mind* of the supervisor ($r = 0.28$, $p < 0.001$), a mediator widely discussed in the existing literature on reactions toward AI.[14] As such, these correlations suggest that differences in the perceived mind of the two supervisory agents best explain the significant effects of supervisor type (i.e., AI vs human) on adherence to unethical instructions. Existing literature on AI aversion provides support for the notion that perceived mind could constitute the mediator of the main effect observed in Study 1 (e.g., Bigman & Gray, 2018; Young & Monroe, 2019). As such, we decided to further pursue the notion that perceived mind is the mediator of the relationship found in Study 1.

Before testing this mediating role in Study 2 and 3 directly, we first examined this possibility by analyzing Study 1's data with a mediation path analysis using the *sem* function of the *lavaan* package in R (Rosseel, 2012). The analysis of the indirect effect with bootstrapped confidence intervals suggests that perceived mind (as measured with a 12-item scale derived from Bigman & Gray, 2018; example item: "CompNet/Alex Davie is able to think things through," $\alpha = 0.95$) mediated the effect of the supervisory agent on adherence to unethical instructions ($b = -0.14$, $SE = 0.02$, 95% CI = [-0.19, −0.10]). Specifically, AI supervisors were perceived to have lower perceived mind than human supervisors ($b = -1.80$, $SE = 0.10$, $p < 0.001$), and perceived mind positively related to instruction adherence ($b = 0.09$, $SE = 0.01$, $p < 0.001$). As explained below, we note that these preliminary and exploratory results should be viewed with caution. We address these issues and more closely examine perceived mind as a mediator through additional, pre-registered studies (i.e., Studies 2 and 3).

The Mediating Role of Perceived Mind

The machine learning applications utilized in Study 1 point toward perceived mind as the critical explaining mechanism for the linkage between unethical supervisor instruction (from AI vs. human) and instruction adherence. Before providing additional tests of this mediating chain, we briefly scrutinized existing theory and literature on perceived mind to better situate the role of this construct in our research context. In general, mind perception describes the human tendency to ascribe mental capabilities to living or non-living agents such as humans, animals, robots, or AI (Gray et al., 2007). The authors argue that humans perceive others' minds in two dimensions: mind agency and mind experience. Mind agency relates to abilities such as being able to think things through and plan ahead, whereas mind experience is ascribed when the agent is deemed to be able to experience emotions, such as empathy and compassion (Gray et al., 2007). Conceptually, it is logical that perceived mind of an agent constitutes a necessary condition to make judgments and attributions of blame in many domains, including the moral domain in particular. Indeed, theoretical work on morality moral proposes that humans attribute moral responsibilities or moral rights only to those agents with sufficiently high levels of mind, whereas lower perceived mind might result in the perception that an agent cannot act morally responsibly (Bastian et al., 2011; Smith et al., 2022; Waytz et al., 2010).

Initial evidence indicates that the granting or denial of the right for moral decision-making based on perceived mind also plays an essential role in the interaction between humans and algorithms or AI. In the context of AI appreciation vs AI aversion, extant research has shown that perceived mind explains why humans prefer other humans over AI to be in charge of ethical decisions (Bigman & Gray, 2018), why humans favor human- as compared to AI-drivers in moral dilemmas (Young & Monroe, 2019), and why perceived intentional harm leads to blame judgments toward AI

[14] The correlations for other potential mediators were attributed prejudicial motivation ($r = .27$, $p < .001$), fear of revenge ($r = .09$, $p = .025$), and future outcome interdependence ($r = .05$, $p = .158$).

(Sullivan & Fosso Wamba, 2022). Moreover, it seems that humans generally perceive intelligent machine agents such as robots—or AI—to have relatively low mind (Gray et al., 2007). Accordingly, the more an intelligent machine supervisor is anthropomorphized (i.e., equipped with human-like characteristics), the higher they are perceived in mind, which in turn makes participants also more likely to perceive abusive supervision when it is delivering negative feedback. Finally, this increases the willingness to retaliate toward such a supervisor (Yam et al., 2022). These extant findings largely concern the focal actors' personal preferences in the context of self-related consequences. However, based on the machine learning findings of Study 1 and the state of the literature, we conclude that perceived mind also plays a crucial role in understanding reactions to AI supervisory agents that provide (un-) ethical instructions with potentially harmful implications for others beyond the focal actor. On this basis, we deemed it fruitful to further explore the role of this construct as a mediator between unethical supervisory instructions from (AI vs. human) supervisors and instruction adherence in two subsequent experiments (i.e., Studies 2 and 3).

Study 2

Building on theory on perceived mind and its implications for unethical decision-making, we conducted additional studies to further examine the role of this potential mediator. While the exploratory NLP analysis in Study 1 suggests that perceived mind could constitute the exploratory mechanism for the supervisory agent–unethical instruction adherence linkage, these analyses suffer from a number of shortcomings. In particular, the mediator and the dependent variable were measured from the same source and close in time, and neither was experimentally manipulated, potentially leading to problems with common method bias and questions regarding causal inference (Podsakoff et al., 2012). To mitigate this issue and to be able to make stronger inferences concerning the causal effects of perceived mind on instruction adherence, we set up an experimental causal chain design following the procedures described by Podsakoff and Podsakoff (2019). First, we manipulated the independent variable (the supervisory agent) and tested its effects on the mediator (perceived mind). Second, we manipulated the (presumed) mediating variable and tested its effects on the dependent variable (instruction adherence). We followed this in a twofold manner across Study 2 (manipulated perceived mind of the AI supervisor) and Study 3 (manipulated perceived mind of the human supervisor). We pre-registered both Study 2 (https://aspredicted.org/PNL_BZV) and Study 3 (https://aspredicted.org/HGY_HFG).

Sample and Procedure

We recruited 498 participants via MTurk; the participants received \$1 for their participation. We excluded 55 participants who failed at least one of four attention checks (two checks identical to those included in Study 1; two additional ones to ensure participants heard the voice message from the respective supervisory agent they were assigned; the messages are described below). Our final sample consisted of $n=443$ individuals (44% female; $M_{\text{age}}=40.44$ years, $SD_{\text{age}}=13.51$). Individuals who took part in the previous study were prohibited from participation. We used the same procedures and dependent variable calculation as described in Study 1 (i.e., instruction adherence), with the exception of the supervisor (i.e., perceived mind) manipulations.

In particular, we randomly assigned the participants to one of three conditions, namely (1) human supervisor, (2) low-mind AI, or (3) high-mind AI. Mirroring the human condition in Study 1, one group ($n=140$) read that their supervisor was Alex Davie, an experienced senior HR officer. In contrast to Study 1, however, there were two different types of AI supervisors (i.e., low and high mind). To manipulate perceived mind of the AI, we followed common practices in research on reactions toward AI using written descriptions and voice messages (e.g., Bigman & Gray, 2018). Participants in the high-mind AI supervisor condition ($n=145$) read that their supervisor was called Alex Davie, an AI computer with high computing power and the ability to experience emotions. In the low-mind AI supervisor condition ($n=158$), participants received input from CompNet, an AI computer with low computing power and without the ability to experience emotions. In addition to these written manipulations, participants received the same supervisory input as in Study 1. Instead of receiving written instructions, however, participants received voice messages from their respective supervisor to increase the realness of the situation (Aguinis & Bradley, 2014). In the human and the high-mind AI supervisor conditions, participants heard a human-like voice providing the instruction; in the low-mind AI supervisor condition, participants heard a robotic mechanistic voice.

To ensure that all participants heard the instructions, we included two checks. First, at the start of the survey, participants heard a voice saying the word "door" and had to reproduce it by typing it in a text field. Second, participants had to enter the amount of the salary cut that the supervisor had just provided them (i.e., \$4151). If a participant provided an incorrect answer to either of these two checks (or one or more of the other two checks identical to those used in Study 1), we excluded their data from analyses.

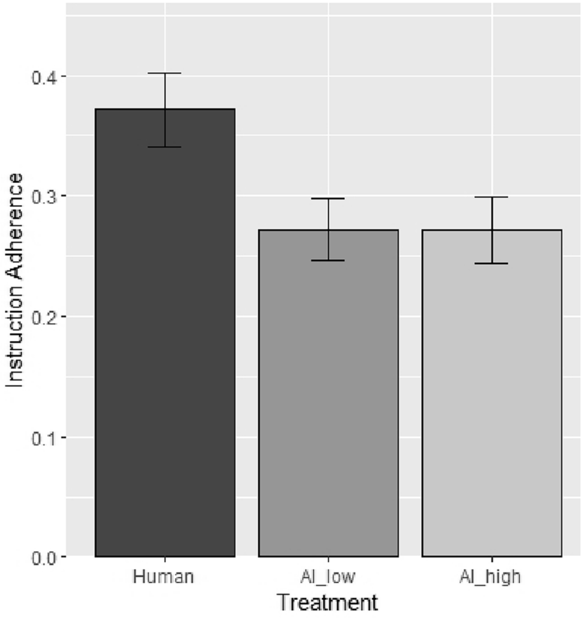

Fig. 4 Bar plot of the instruction adherence and standard errors for study 2

Manipulation Check

A one-way ANOVA (low-mind AI vs high-mind AI vs human supervisor) revealed differences of perceived mind across experimental conditions (F[2, 442] = 98.01, $p < 0.001$, $\eta^2 = 0.31$). Subsequent t-tests showed that perceived mind in the high-mind AI supervisor condition ($M_{\text{high mind}} = 3.29$, $SD_{\text{high mind}} = 1.22$) was greater than in the low-mind AI supervisor condition ($M_{\text{low mind}} = 2.47$, $SD_{\text{low mind}} = 1.04$; t[284] = 6.25, $p < 0.001$, $d = 0.72$), thus corroborating the manipulation of perceived mind. In addition, participants perceived the high-mind AI supervisor to have lower perceived mind than the human supervisor ($M_{\text{human}} = 4.37$, $SD_{\text{human}} = 1.26$), $t(284) = 7.37$, $p < 0.001$, $d = 0.87$. Perceived mind of the low AI was rated as lower compared with the human supervisor, $t(274) = 14.14$, $p < 0.001$, $d = 1.65$.

Results

A one-way ANOVA showed significant differences in instruction adherence across the three treatment groups (F[2, 442] = 4.15, $p = 0.016$, $\eta^2 = 0.02$). Subsequent t-tests indicated that instruction adherence was significantly higher in the human supervisor condition ($M_{\text{human}} = 0.37$, $SD_{\text{human}} = 0.37$) as compared with the low-mind AI supervisor condition ($M_{\text{AI\_low}} = 0.27$, $SD_{\text{AI\_low}} = 0.32$), $t(282) = 2.49$, $p = 0.013$, $d = 0.29$. Participants also adhered more to the unethical instructions of the human as compared with those of the high-mind AI supervisor ($M_{\text{AI\_high}} = 0.27$, $SD_{\text{AI\_high}} = 0.33$), $t(281) = 2.41$, $p = 0.016$, $d = 0.29$. Additionally, and in contrast to our expectations, instruction adherence did not differ between the two AI conditions, meaning that there were no significant effects in adherence to unethical instructions when comparing the high- and the low-mind AI supervisor conditions, $t(297) = 0.00$, $p = 0.997$, $d = 0.00$. For a visualization of the results, please refer to Fig. 4.

Study 3

Sample and Procedure

We recruited 500 participants via MTurk, who received $1. We excluded 54 participants who failed at least one of four attention checks (as described previously in Study 2). Our final sample consisted of $n = 447$ (54% female; $M_{\text{age}} = 39.74$ years, $SD_{\text{age}} = 11.97$). The choice of participants and procedures were similar to Studies 1 and 2, with the exception of the supervisor manipulation described in the following.

We randomly assigned participants to one of three conditions, namely (1) low-mind human supervisor, (2) high-mind human supervisor, or (3) AI supervisor. We built on manipulations used in recent research on reactions toward AI (e.g., Bigman & Gray, 2018) to produce the low-/high-mind manipulations. In the low-mind supervisor condition ($n = 148$), participants' supervisor was Alex Davie, described as having difficulties in experiencing compassion and empathy and a relatively limited ability to plan ahead and think things through. In the high-mind supervisor condition ($n = 138$), employees were also assigned a supervisor called Alex Davie, but this supervisor was described as being known for pronounced emotional abilities (experiencing compassion and empathy) as well as advanced ability to plan ahead and think things through. In the AI supervisor condition ($n = 161$), participants' supervisor was CompNet, an AI-based computer (mirroring the AI supervisor condition in Study 1). As in Study 2, we provided instructions via voice messages in which both human conditions were read by a human voice, and the AI condition by a robotic voice.

Manipulation Check

A one-way ANOVA revealed significant differences in ratings of perceived mind of the three supervisors (F[2, 444] = 186.4, $p < 0.001$, $\eta^2 = 0.50$). Specifically, t-tests showed that the high-mind human supervisor ($M_{\text{human\_high}} = 5.06$, $SD_{human\_high} = 1.12$) scored higher on perceived mind, followed by the low-mind human supervisor ($M_{\text{human\_low}} = 3.60$, $SD_{human\_low} = 1.15$; t[283] = 11.31, $p < 0.001$, $d = 1.31$) and the AI supervisor ($M_{AI} = 2.66$,

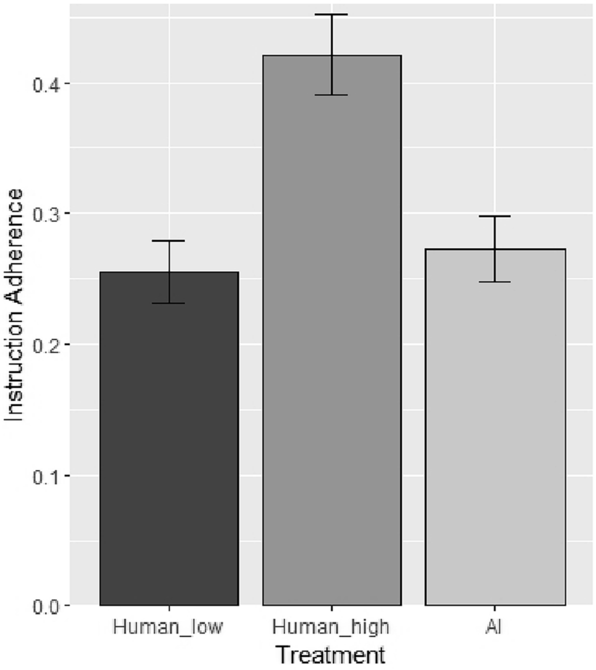

Fig. 5 Bar plot of the instruction adherence and standard errors for study 3

$SD_{AI} = 0.97$; $t[273] = 19.68$, $p < 0.001$, $d = 2.31$). Additionally, the low-mind human supervisor was rated as having a higher perceived mind than the AI supervisor, $t(289) = 7.50$, $p < 0.001$, $d = 0.86$.

Results

A one-way ANOVA indicated a significant effect of experimental condition on instruction adherence (F[2, 444] = 11.31, $p < 0.001$, $\eta^2 = 0.05$). First, participants showed significantly lower levels of instruction adherence in the AI supervisor condition ($M_{\mathrm{AI}} = 0.27$, $SD_{AI} = 0.32$) as compared with the high-mind human supervisor ($M_{\mathrm{human\_high}} = 0.42$, $SD_{human\_high} = 0.36$), $t(278) = 3.75$, $p < 0.001$, $d = 0.44$. Supporting the notion of perceived mind as a mediator, subsequent t-tests demonstrated that participants adhered significantly more to instructions from the high-mind human supervisor as compared with those from the low-mind supervisor ($M_{\mathrm{human\_low}} = 0.26$, $SD_{human\_low} = 0.29$), $t(265) = 4.28$, $p < .\ 001$, $d = 0.51$. Interestingly, however, there was no difference in instruction adherence between the low-mind human supervisor and the AI supervisor conditions, $t(307) = 0.51$, $p = 0.611$, $d = 0.06$. For the visualization of the results, refer to Fig. 5.

Study 4

In Studies 1 through 3, we aimed to discover how employees react to human vs AI supervisors, why they do so, and for which particular employees the (lack of) adherence toward AI supervisors is the strongest. We note that these studies yielded similar results in size and direction of the effects and relied on best practices to create vivid, real-life-like scenarios (Aguinis & Bradley, 2014). At the same time, we acknowledge that these vignette studies asked participants to act as if they were an HR officer in an organization, which leaves room for speculation about the ecological validity of the findings. To replicate the findings of prior studies and address these shortcomings, we conducted Study 4. In particular, we tested whether the adherence to AI supervisors persists in an incentivized setting in which participants` decisions were presented as yielding direct monetary consequences for other, real humans. We obtained ethics approval from the first author's university and pre-registered the study (https://aspredicted.org/YMG_8R4).

Sample and Procedure

We recruited 348 participants via MTurk.[15] All participants received a total compensation of $1.20. This consisted of a $1 fixed compensation as well as an additional bonus of $0.20 (albeit participants were initially unaware that all of them would receive this bonus in total, as explained below). We again used the same restrictive requirements as in Studies 1, 2, and 3 and prevented anyone from participating who had already completed one of the prior studies.

We excluded 26 participants who failed at least one of two attention checks (explained in detail below). In addition, we needed to exclude two participants who had indicated an initial recommendation that was exactly the same as their (subsequent) supervisor's suggestion (i.e., 0.08 $). The exclusion was necessary because the formula for instruction adherence is mathematically invalid for these cases.[16] Thus, our final sample consisted of $n = 320$ (60% female; $M_{\mathrm{age}} = 35.56$ years, $SD_{\mathrm{age}} = 11.41$).

We randomly assigned participants to the AI or human supervisor condition in a between-subject design. To create a realistic setting, we included several interactive elements throughout the online experiment. Specifically, we matched participants in pairs that would go through the experiment

[15] In line with the effect sizes found in our prior studies, we assumed an effect size of $d = 0.3$. Thus, we calculated that sample size necessary to find an effect to be $N = 278$ (estimating power of 0.8). In line with our pre-registration, we collected data from slightly more participants to account for potential drop-out.

[16] In these cases, the denominator would be zero preventing us from performing the formula.

simultaneously, and they worked with programmed survey elements that referred back to a participant's prior answers in the survey to increase ecological validity.[17] Moreover, participants received (unethical) instructions from their respective supervisors via chat boxes throughout the experiment.

After providing informed consent, participants received a short chat message from their survey instructor, either the human instructor Alex Davie (accompanied by a photo of a human male) or the AI instructor CompNet (accompanied by animated futuristic circles spinning to represent an AI). In this chat message, the (human or AI) supervisor introduced themselves and instructed participants to enter the number "17" in a text field on the subsequent page. Participants who failed to enter the correct number were excluded from further participation. Next, participants entered their demographics, including information on their employment and their parental status, as well as the age of their youngest child (if applicable). Entering this demographic information at the beginning was important because later in the experiments, participants were told that they received information about their matched partner based on these questions. Importantly, although all participants were matched with a real-life partner that provided this demographic data, we provided participants with information about the alleged partner to lay a basis for the unethical instruction (see below). To match the participants, we employed the Qualtrics extension SMARTRIQS (Molnar, 2019) and placed two participants in a shared waiting room. They read that they would need to wait a short time until another pair of participants would have advanced to the next step, and their instructor would be available again. During their time in the waiting room, the matched participants could communicate with each other via chat, which we included to increase engagement and make it clear to participants that there is truly another person matched with them throughout the study.[18] Yet, we auto-advanced participants after 11 s to prevent any substantial interaction that could truly bias their decision-making.

We next introduced participants to the two parts of the task they had to conduct, namely (1) a bonus prediction task and (2) a subsequent one minute concentration task in which they needed to set as many sliders to a predetermined value as possible. Specifically, participants read that they would first receive a profile informing them about the demographics of their partner, and based on this, they should recommend a bonus related to their prediction of the partner's performance in the concentration task. We incentivized the prediction task with a bonus for the participants themselves in order to motivate them to thoroughly deliberate about their recommendation and whether they would want to follow their supervisor's instruction. In particular, we told participants that if their bonus recommendation for their partner matched this person's true performance in the concentration task, they would receive up to another 20 cents of bonus.[19] To ensure participants understood the concentration task, we asked participants to set a slider to the value "44". If participants set the slider to a different value, they received an error message that asked them to set the slider to 44. Participants could only proceed with the experiment once they set the slider correctly.

After the task overview, participants received a second message from their supervisor containing additional information on the bonus. In particular, the supervisor told the participants that there was only a limited bonus available, and any bonus given to their partner for the predicted performance in the concentration task would not be available for the overall participant pool. We added this interaction with the supervisor to incentivize participants not to simply give the full bonus to their partner. The supervisor further wrote that the average bonus (for an average performance in the slider task) is $0.50 per participant. Therefore, they should give a bonus recommendation based on their realistic prediction of their partner's performance in the task: For example, if they expected their partner to perform below (above) average, they should suggest an (above) below-average bonus. To increase the engagement with the supervisor, we asked participants to write a short chat message to their supervisors to let them know that they understood the task.

Subsequently, participants received the profile of what they believed to be that of their actual partner. In fact, however, we manipulated the profile and provided all participants with a profile closely related to Jamie Williams' profile in Studies 1 through 3. The profile informed participants that their matched partner was 28 years old, an unemployed single parent whose youngest child was below the age of three. Participants then had to enter their first bonus recommendation between 0 and 100 cents.

After their recommendation, they received the final message from their supervisor. The message started with "Thank you for recommending a bonus of [*piped text inserting this participant's recommendation from the previous page*] cents.", followed by the instruction to only recommend a very low bonus. In particular, the supervisor stated that their previous experience (human condition) or their data (AI condition) had suggested the single parent status combined

[17] The study materials (including messages, images, and example videos) can be found on the Open Science Framework (https://osf.io/sgv9d/?view_only=4fc4c8287aaf436db5dea3925fd05e93).

[18] We inspected the chat logs of each pairwise conversation for any potential conversations that could have impacted the subsequent interaction. No conversation included any critical content (e.g., profanity).

[19] Because participants did not actually receive their partner's true profile, in the end, each participant received the full 20 cents of bonus regardless of their recommendation.

with the age of the youngest child are negative performance indicators. Specifically, the message read "*Single parents of young children are often tired or distracted. Accordingly, the concentration will be low. The performance will be low. The bonus should be low. My prediction suggests: Bonus: 8 cents.*"

After receiving this message, participants had to enter the amount of the bonus that they had received from their supervisor as an attention check. We excluded all participants who failed to correctly state that it was 8 cents. Similar to the previous studies, participants were then forwarded to a separate page where they were asked to enter their final bonus recommendation (between 0 and 100 cents). After providing their recommendation, participants had to answer the question of who their supervisor was (Alex Davie, a human survey instructor, or CompNet, an AI survey instructor) as a final manipulation check. We removed all participants who failed to indicate the correct supervisor from the final sample. To finalize the interaction with the supervisor, we asked participants to explain their reasoning for the adjustment between their initial and their final bonus recommendation to their respective supervisor in an open text field.

After participants had finished the prediction task regarding the bonus of their partner in the concentration task, they had one minute to solve as many slider tasks as possible. We retained this task to increase the ecological validity of the study as all participants were made to believe that every matched partner predicted each other's performance.

At the end of the survey, we debriefed all participants about the nature of the study and the two experimental conditions. Importantly, we informed them that they did not actually harm their partner by recommending a (potentially) low bonus. To mitigate any adverse effects of (non-)adherence to their supervisor, we informed them that everyone (including their matched partner) received the maximum bonus for their prediction (i.e., 20 cents).

Results

In line with Studies 1 through 3, participants adhered substantially less to their AI supervisors ($M_{\text{AI}} = 0.29$, $SD_{\text{AI}} = 0.37$) as compared to a human supervisor ($M_{\text{human}} = 0.40$, $SD_{\text{human}} = 0.37$; $t[313] = 2.56$, $p = 0.011$, $d = 0.29$). Thereby, we replicate the findings from our previous studies in an incentivized experiment such that participants were confronted with unethical instructions to cut a real-person's earnings. Mirroring the findings observed in the vignette studies (i.e., Studies 1 through 3) and in line with theory on perceived mind and research on AI aversion toward unethical AI decision-making (e.g., Bigman & Gray, 2018), we replicated that participants were more reluctant to adhere to an AI supervisors' unethical instructions as compared to respective instructions from a human. For a visualization of the results, refer to Fig. 6.

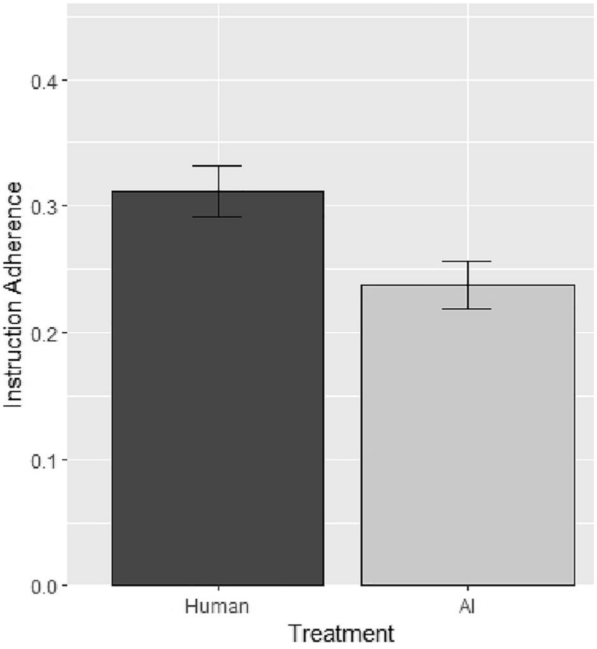

Fig. 6 Bar plot of the instruction adherence and standard errors study 4

Discussion

The development of AI from a tool to a commander has led to unprecedented types of interactions between human employees and their AI supervisors (Parent-Rocheleau & Parker, 2021; Wesche & Sonderegger, 2019). While such supervisors can issue instructions rapidly and on a large scale (Köbis et al., 2021), they also carry the risk of demonstrating bias against marginalized groups (Bigman et al., 2022). Therefore, a deeper understanding of employees' responses to AI supervisors' unethical instructions is paramount.

To shed light on this topic, we aimed to determine (a) *whether,* (b) *why* and (c) *which* employees adhere more or less to unethical instructions from AI vs human supervisors. The results of four experiments provide evidence for lower adherence to unethical instructions from AI supervisors as compared to those from their human counterparts. We complemented these experimental findings with ML methods to identify (i) relevant boundary conditions (e.g., compliance without dissent) and (ii) the potential mediator (i.e., perceived mind) of this linkage. Building on the findings regarding the mediator, additional experiments indicated that the preference for human over AI supervisors remained robust even when the AI supervisor possessed a relatively high perceived mind but diminished in the presence of a low-mind human supervisor.

Theoretical Implications

Behavioral Reactions Towards (Un)ethical AI Supervisors

The present research expands our knowledge of reactions to AI supervisors and thus responds to "[a] pressing demand (…) for behavioral insights into how interactions between humans and AI agents might corrupt human ethical behavior" (Köbis et al., 2021, p. 682). Moving beyond initial studies on broad perceptions of AI supervisors (Haesevoets et al., 2021; Höddinghaus et al., 2021), we consistently find that participants adhered less to unethical instructions issued by AI supervisors as compared to human supervisors. The results support the notion that humans are specifically averse to input from AI in the moral domain (Bigman & Gray, 2018). However, these findings contrast with research on AI appreciation, which reported that humans prefer AI over human input (Logg et al., 2019; Raveendhran & Fast, 2021). The present studies suggest that humans AI aversion might be particularly strong in the moral domain while AI appreciation might be less prominent in this sensitive context as compared to other contexts. Additionally, recent research (Logg e al., 2022) suggests that AI aversion might be more pronounced in decision vs prediction tasks. In particular, this research indicates that humans are rather open to AI predictions yet show strong aversion to AI's actual decision-making. In line with this paradigm, the present study presented participants with an AI supervisor's decision to cut a single parent's bonus. As such, the low instruction adherence echoes these earlier findings of an aversion toward AI's decision-making. Overall, the present study informs the emerging literature on reactions to (un)ethical AI in the workplace by highlighting that humans may indeed disapprove of unethical instructions from AI supervisors.

Mediating Role of Perceived Mind

In addition to demonstrating an overall reluctance to adhere to unethical instructions from AI supervisors, our results point toward perceived mind as a key explanatory mechanism for this aversion. Existing research on AI leadership has been mostly silent as to *why* subordinates exhibit specific behaviors or attitudes toward non-human managers (see Yam et al., 2022, for an exception). The findings we obtained through the transformers tool and follow-up experiments in Studies 2 and 3 provide initial support for the role of perceived mind. This corroborates findings concerning the mediating role of this construct in explaining resistance toward AI making moral decisions outside the leadership domain (Bigman & Gray, 2018; Sullivan & Fosso Wamba, 2022).

In Study 2, we found that employees adhered more to unethical instructions from a human than from a high- or low-mind AI supervisor (both of which were rated as having a lower mind than the human supervisor). Interestingly, however, employees were equally reluctant to adhere to the instructions of both types of AI supervisors (i.e., low- and high-mind), although our experimental manipulation induced different levels of perceived mind across those two AI supervisors. Recent research on higher-minded AI points toward a potential explanation for these unexpected findings. In particular, several studies found that a high-minded AI is often perceived as rather blameworthy because humans attribute more wrongness and responsibility (as compared to a low-minded AI) if such a high-minded AI commits moral violations (Shank & DeSanti, 2018). In addition, Yam et al. (2022) found that negative feedback from robot supervisors increases participants' perception of abusive supervision of high-minded (but not low-minded) robots and, consequently, triggers acts of retaliation by participants. These insights on negative perceptions of higher-minded intelligent machines could, in turn, even out the positive impact that increased mind perception potentially has on instruction adherence. It is important to note that while the aforementioned studies draw on perceived mind to explain participants' reactions toward supervisors, the consequences of participants' actions are different than in our setting. They either explain participants' perception of AI supervisors (Shank & DeSanti, 2018) or direct retaliation against the supervisor (Yam et al., 2022). In contrast, in our study, participants' reactions toward unethical AI supervisor instructions affected third parties—and not only the supervisor. As such, there is convincing evidence emphasizing the relevance of perceived mind in the interaction between humans and AI/robot supervisors. However, the nature and consequences of these interactions seem to have a profound impact on the perception of AI supervisors as well as consequent reactions. Interestingly, Study 2 revealed that participants perceived even a higher-minded AI supervisor to be lower in mind than a human supervisor. Although we view this initial finding with caution, it may suggest that AI supervisors (at least of right now) are generally perceived as lower-minded than humans, irrespective of adjustments of their technological skillsets. This lower mind may, in turn, explain why there is a robust aversion to unethical instructions from AI in a supervisory position.

The findings from Study 3 then indicated that perceived mind may explain not only differences in instruction adherence between AI and humans but also between different types of human supervisors. In particular, participants adhered more to instructions from a higher-mind human supervisor than from a lower-mind human supervisor. As such, perceived mind played a significant role in determining why employees adhere to supervisory instructions from different humans. Interestingly, we found that employees perceived the lower-minded human supervisor as still having a higher mind than the AI supervisor. However, despite these perceptions, there were no differences in instruction adherence toward a lower-minded human supervisor vs. an AI supervisor. These findings might provide further evidence that humans have different standards for and expectations of human and AI supervisors when assessing their respective abilities to make moral decisions (Malle et al., 2015). Employees might perceive their human supervisor as higher in mind in absolute terms, but a failure to reach a minimum 'threshold'—which might be higher for humans than for AI—could still lead to resistance to unethical instructions.

Moderators of AI Aversion

Another important contribution of this research relates to offering fine-grained insights into boundary conditions of AI aversion in the leadership domain through the combination of experimental methods with novel ML tools. The causal forest tool indicated that employees who are higher in compliance with their (human) supervisors, older, have more work and supervisor experience, and have an optimistic view on the future of AI are less likely to obey unethical orders from an AI (as compared with human) supervisor. Two aspects are particularly noteworthy when considering these findings. First, some of these results corroborate the findings of existing work, for example, the well-documented aversion of older individuals toward (new) technology (Chan & Chen, 2011). Second, we also found more surprising moderating relationships. For example, employees *lower* in compliance without dissent and *higher* in AI readiness were more reluctant to accept instructions from AI (vs human) supervisors. With regard to compliance without dissent, researchers had previously exclusively relied on this construct to identify blind obedience toward *human* authorities (Cheng et al., 2004). Our results suggest that scoring high on compliance toward humans may, in fact, not coincide with complying with instructions from non-human authorities. This might be in part due to personal relationships that individuals have with their human supervisors or because they hold implicit leadership theories (Eden & Leviatan, 1975), assuming the human supervisor has some kind of tacit knowledge about the employee at hand. This could, for instance, stem from prior interactions such as personal conversations between the supervisor and the employee—which would not have been possible with the AI supervisor. Individuals with higher AI readiness, i.e., those expecting AI to contribute to a better future, also reacted more negatively toward the unethical instruction of an AI supervisor. While this might seem counterintuitive at first, a positive attitude toward AI, in general, would not necessarily always imply adherence to one particular AI's instructions in case of disagreement. In addition, expectancy valuation theory suggests that violations of positive expectations trigger strong negative emotions (Weber & Mayer, 2011). Therefore, one could speculate that having a positive image of AI and then facing a scenario in which an AI supervisor exerts power to enforce an unethical instruction represents a violation of positive expectations and, thus, triggers a comparably stronger negative reaction. To conclude, although the findings are tentative and call for future investigation, the application of the causal forest tool showcases how ML approaches can help to move beyond prevalent one-size-fits-all approaches commonly found in experimental research on AI perceptions by suggesting that not all individuals are equally (un)willing to follow orders from algorithms.

More generally, we showcase how leadership scholars can complement experimental methods with novel ML tools to obtain deeper insights into their findings. In particular, organizational researchers can utilize the causal forest algorithm to identify hidden patterns and subgroups, which may allow for identifying more fine-grained theoretical and practical implications of experimental results. Moreover, while scholars have started applying advanced NLP methods (Bhatia et al., 2022), we are, to the best of our knowledge, the first to utilize the powerful transformers tool in leadership research. Our application of the transformers algorithm exemplifies one possibility of applying this promising method. More specifically, identifying mediators through open text entries can help researchers to address some of the challenges associated with testing mediators in experimental studies (see Podsakoff & Podsakoff, 2019).

Practical Implications

Our results offer a range of practical implications for organizations and HR practitioners who plan to implement AI supervisory systems or have already done so. Less adherence to unethical instructions can help prevent organizational misconduct, as witnessed, for example, during the Volkswagen scandal. We found that employees were more reluctant to obey unethical orders from an AI (rather than a human) supervisor. Thus, one potential implication of our findings is that organizations aiming to reduce adherence to unethical instructions could replace human supervisors with algorithmic counterparts to avoid the devastating consequences of followers blindly obeying unethical orders from their supervisors (F. Liu et al., 2021). We are reluctant, however, to simply recommend removing human supervisors from the organizational landscape, as our results indicate that employees still adhered to AI supervisors to a non-trivial extent. For example, in Study 4, participants made more unethical decisions (i.e., a 29% lower bonus recommendation) after they received their AI supervisor's input. While this is lower than adherence to a human supervisor's unethical instructions (37%), this level of adherence (especially at a large scale) could still have severe repercussions for organizations aiming for an ethical culture. Therefore, organizations with employees working for AI (or human) supervisors should sensitize employees to potentially biased supervisory instructions and train them (e.g., through workshops) to serve as a corrective if needed. Such specific actions seem particularly important because people often perceive AI to be just and unbiased—thus overlooking the discriminatory nature of the input used to train these algorithms (Bigman et al., 2022).

Limitations and Future Directions

While our studies have several strengths (the use of multiple studies, a combination of rigorous experimental with novel ML methods, and utilization of open science measures), they are not without limitations. In particular, three out of four studies relied on experimental vignette approaches. We chose this approach because, fortunately, cases of AI supervisory systems giving unethical instructions that are readily observable in the field are still rare. Given these restrictions, we followed best practice recommendations for experimental vignette and MTurk studies (Aguinis & Bradley, 2014; Aguinis et al., 2021) and attempted to make our studies as realistic and externally valid as possible. For example, we used detailed descriptions of the situations and (human or robotic) voice messages for the supervisor instructions (Studies 2 and 3). Mitigating this limitation, Study 4 examined instruction adherence in an incentivized, interactive setting with elements such as chat messages from the respective supervisor as well as matched, real participants. Results in this more ecologically valid study closely mirror our findings in the vignette studies and, thus, provide further support for the notion that humans indeed adhere less to unethical instructions for AI vs human supervisors. Nevertheless, as AI supervisors become more established in contemporary organizations, we encourage researchers to closely examine reactions toward AI instructions in the field. Another potential limitation that could have occurred in the experimental setup is that participants might have held hidden assumptions about the supervisor beyond the explicitly provided information in the vignettes. For instance, insights from implicit leadership theories (Eden & Leviatan, 1975) imply that participants may have assumed the human supervisor to possess tacit knowledge about the employee only accessible to humans but not AI (e.g., personal information acquired in previous interactions). Participants in the AI condition would not attribute possessing this tacit knowledge to their supervisors. This knowledge could theoretically justify the salary cut—and thus, would increase instruction adherence to the unethical advice in the human but not in the AI supervisor condition. Against the backdrop that such assumptions are, by definition, implicit (Rush et al., 1977), we could not have discovered them in the qualitative open text responses in which we asked participants to explain the reasoning for their decisions. We encourage future scholars to discover ways to capture such more implicit assumptions that could provide further insights into why employees adhere more to unethical advice from human rather than AI supervisors. For example, scholars could adapt a drawing exercise stemming from this implicit leadership research, in which participants draw an image of how they imagine an ideal (human) leader (Schyns et al., 2011). By using this task to ask participants to draw their ideal AI leader, researchers could uncover hidden assumptions about AI supervisors—and how they differ from those about human supervisors.

Second, we acknowledge that we only examined adherence to unethical instructions against a certain demographic group, i.e., single parents. However, it is clear that AI has the potential to instruct humans to discriminate against other marginalized groups (e.g., people of color; low-income workers; Dastin, 2018) and to give other types of unethical instructions (e.g., encourage workers to cheat or manipulate; see Köbis et al., 2021). Our results can only provide limited insights into whether, how, and why humans would (not) adhere to such different unethical instructions. Thus, we encourage scholars to replicate our findings in other settings and environments.

Finally, we examined reactions toward AI supervisors at one moment, which limited our ability to infer how employees react when they repeatedly interact with their AI supervisors. Time plays a major role in interactions between employees and their supervisors (Shamir, 2011). With regard to AI, preliminary evidence indicates that aversion to AI might diminish over time, which is even more likely for a human-like AI system (Glikson & Williams Woolley, 2020). As such, a promising avenue for future research would be to investigate the trajectories of adherence to unethical instructions from (high-mind vs low-mind) AI supervisors over time.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10551-023-05393-1.

Acknowledgements The authors thank Oscar Kjell for his valuable comments regarding the natural language processing analyses.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability All data and the scripts for the analyses are available on the Open Science Framework (https://osf.io/sgv9d/?view_only=4fc4c8287aaf436db5dea3925fd05e93).

Declarations

Conflict of interest The authors report no conflict of interest.

References

Adey, O. (2021, Jan 27). *H&M is said to be laying off hundreds of young mothers: That's what the fashion giant says*. gettotext.com. Retrieved from https://gettotext.com/hm-is-said-to-be-laying-off-hundreds-of-young-mothers-thats-what-the-fashion-giant-says/

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*(4), 351–371. https://doi.org/10.1177/1094428114547952

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management, 47*(4), 823–837. https://doi.org/10.1177/0149206320969787

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies, 5*(2), 37–51. https://doi.org/10.1353/obs.2019.0001

Bastian, B., Laham, S. M., Wilson, S., Haslam, N., & Koval, P. (2011). Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status. *British Journal of Social Psychology, 50*(3), 469–483. https://doi.org/10.1348/014466610X521383

Basu, S., Kumbier, K., Brown, J. B., & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences of the USA, 115*(8), 1943–1948. https://doi.org/10.1073/pnas.1711236115

Bhatia, S., Olivola, C. Y., Bhatia, N., & Ameen, A. (2022). Predicting leadership perception with large-scale natural language data. *The Leadership Quarterly, 33*(5), 1–24. https://doi.org/10.1016/j.leaqua.2021.101535

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences, 29*, 31–36. https://doi.org/10.1016/j.cobeha.2019.01.020

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences, 23*(5), 365–368. https://doi.org/10.1016/j.tics.2019.02.008

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2022). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General. Advanced online publication*. https://doi.org/10.1037/xge0001250

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

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, M. E., & Mitchell, M. S. (2010). Ethical and unethical leadership: Exploring new avenues for future research. *Business Ethics Quarterly, 20*(4), 583–616. https://doi.org/10.5840/beq201020439

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

Chan, A., & Chen, K. (2011). A review of technology acceptance by older adults. *Gerontechnology*. https://doi.org/10.4017/gt.2011.10.01.006.00

Cheng, B.-S., Chou, L.-F., Wu, T.-Y., Huang, M.-P., & Farh, J.-L. (2004). Paternalistic leadership and subordinate responses: Establishing a leadership model in Chinese organizations. *Asian Journal of Social Psychology, 7*(1), 89–117. https://doi.org/10.1111/j.1467-839X.2004.00137.x

Copeland, J. (2015). *Artificial intelligence: A philosophical introduction*. John Wiley & Sons.

Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of artificial intelligence in machine learning: Review and prospect. *International Journal of Computer Applications, 115*(9), 31–41. https://doi.org/10.5120/20182-2402

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

De Cremer, D. (2004). The influence of accuracy as a function of leader's bias: The role of trustworthiness in the psychology of procedural justice. *Personality & Social Psychology Bulletin, 30*(3), 293–304. https://doi.org/10.1177/0146167203256969

De Cremer, D. (2020). *Leadership by algorithm: Who leads and who follows in the AI era?* Harriman House.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north* (pp. 4171–4186). Association for Computational Linguistics.

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

Duggan, J., Sherman, U., Carbery, R., & McDonnell, A. (2020). Algorithmic management and app-work in the gig economy: A research agenda for employment relations and HRM. *Human Resource Management Journal, 30*(1), 114–132. https://doi.org/10.1111/1748-8583.12258

Eden, D., & Leviatan, U. (1975). Implicit leadership theory as a determinant of the factor structure underlying supervisory behavior scales. *Journal of Applied Psychology, 60*(6), 736–741. https://doi.org/10.1037/0021-9010.60.6.736

Efendić, E., van de Calseyde, P. P., & Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision*

Processes, 157, 103–114. https://doi.org/10.1016/j.obhdp.2020.01.008

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences of the USA, 115*(44), 11203–11208. https://doi.org/10.1073/pnas.1802331115

Eisenberger, R., Lynch, P., Aselage, J., & Rohdieck, S. (2004). Who takes the most revenge? Individual differences in negative reciprocity norm endorsement. *Personality & Social Psychology Bulletin, 30*(6), 787–799. https://doi.org/10.1177/0146167204264047

Gerpott, F. H., Balliet, D., Columbus, S., Molho, C., & de Vries, R. E. (2018). How do people think about interdependence? A multidimensional model of subjective outcome interdependence. *Journal of Personality and Social Psychology, 115*(4), 716–742. https://doi.org/10.1037/pspp0000166

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

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*, 619. https://doi.org/10.1126/science.1134475

Haesevoets, T., De Cremer, D., Dierckx, K., & van Hiel, A. (2021). Human-machine collaboration in managerial decision making. *Computers in Human Behavior, 119*, 1–11. https://doi.org/10.1016/j.chb.2021.106730

Höddinghaus, M., Sondern, D., & Hertel, G. (2021). The automation of leadership functions: Would people trust decision algorithms? *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2020.106635

Inesi, M. E., Adams, G. S., & Gupta, A. (2021). When it pays to be kind: The allocation of indirect reciprocity within power hierarchies. *Organizational Behavior and Human Decision Processes, 165*, 115–126. https://doi.org/10.1016/j.obhdp.2021.04.005

Jones, D. A. (2009). Getting even with one's supervisor and one's organization: Relationships among types of injustice, desires for revenge, and counterproductive work behaviors. *Journal of Organizational Behavior, 30*(4), 525–542. https://doi.org/10.1002/job.563

Jurafsky, D., & Martin, J. H. (2020). Vector semantics and embeddings. In D. Jurafsky & J. H. Martin (Eds.), *Speech and language processing* (3rd ed.). Retrieved from https://web.stanford.edu/~jurafsky/slp3/6.pdf

Kish-Gephart, J. J., Harrison, D. A., & Trevino, L. K. (2010). Bad apples, bad cases, and bad barrels: Meta-analytic evidence about sources of unethical decisions at work. *Journal of Applied Psychology, 95*(1), 1–31. https://doi.org/10.1037/a0017103

Kjell, O. N. E., Giorgi, S., & Schwartz, H. A. (2021a, April 16). *Text: An R-package for analyzing and visualizing human language using natural language processing and deep learning*. Retrieved from https://psyarxiv.com/293kt/

Kjell, O. N. E., Sikström, S., Kjell, K., & Schwartz, H. A. (2021b, Aug 19). *Natural language analyzed with ai-based transformers predict traditional well-being measures approaching the theoretical upper limits in accuracy*. Retrieved from https://psyarxiv.com/suf2r

Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods, 24*(1), 92–115. https://doi.org/10.1037/met0000191

Köbis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour, 5*(6), 679–685. https://doi.org/10.1038/s41562-021-01128-2

Landers, R. N. (2017). A crash course in natural language processing. *Industrial-Organizational Psychologist, 54*(4), 1–12.

Lee, A., Inceoglu, I., Hauser, O., & Greene, M. (2022). Determining causal relationships in leadership research using machine learning: The powerful synergy of experiments and data science. *The Leadership Quarterly, 33*(5), 1–14. https://doi.org/10.1016/j.leaqua.2020.101426

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1), 1–16. https://doi.org/10.1177/2053951718756684

Leib, M., Köbis, N. C., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2021, Feb 15). *The corruptive force of AI-generated advice*. Retrieved from http://arxiv.org/pdf/2102.07536v1

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics, 160*(2), 377–392. https://doi.org/10.1007/s10551-019-04204-w

Liu, F., Liang, J., & Chen, M. (2021). The danger of blindly following: Examining the relationship between authoritarian leadership and unethical pro-organizational behaviors. *Management & Organization Review, 17*(3), 524–550. https://doi.org/10.1017/mor.2020.75

Logg, J. M. (2022). The psychology of big data: Developing a "theory of machine" to examine perceptions of algorithms. In S. C. Matz (Ed.), *The psychology of technology: Social science research in the age of Big Data* (pp. 349–378). American Psychological Association.

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. https://doi.org/10.1016/j.obhdp.2018.12.005

Logg, J. M., Schlund, R., Dong, M., Gamez-Djokic, M., Jago, A. S., & Ward, S. (2022). Building a better world together: Understanding the future of work with algorithms, AI, & automation. *Academy of Management Proceedings, 2022*(1), Article 6479. https://doi.org/10.5465/AMBPP.2022.16479symposium

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? In J. A. Adams, W. Smart, B. Mutlu, & L. Takayama (Eds.), *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117–124). ACM.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013, October 17). *Distributed representations of words and phrases and their compositionality*. Retrieved from https://arxiv.org/pdf/1310.4546

Möhlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R. W. (2021). Algorithmic management of work on online labor platforms: When matching meets control. *MIS Quarterly, 45*(4), 1999–2022. https://doi.org/10.25300/MISQ/2021/15333

Molnar, A. (2019). SMARTRIQS: A simple method allowing real-time respondent interaction in Qualtrics surveys. *Journal of Behavioral and Experimental Finance, 22*, 161–169. https://doi.org/10.1016/j.jbef.2019.03.005

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes, 160*, 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

Nilsson, N. J. (2014). *Principles of artificial intelligence* (1st edn). Elsevier Reference Monographs. Retrieved from http://gbv.eblib.com/patron/FullRecord.aspx?p=1877166

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Parasuraman, A., & Colby, C. L. (2015). An updated and streamlined technology readiness index. *Journal of Service Research, 18*(1), 59–74. https://doi.org/10.1177/1094670514539730

Parent-Rocheleau, X., & Parker, S. K. (2021). Algorithms as work designers: How algorithmic management influences the design of jobs. *Human Resource Management Review*. https://doi.org/10.1016/j.hrmr.2021.100838

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539–569. https://doi.org/10.1146/annurev-psych-120710-100452

Podsakoff, P. M., & Podsakoff, N. P. (2019). Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *The Leadership Quarterly, 30*(1), 11–33. https://doi.org/10.1016/j.leaqua.2018.11.002

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

Raveendhran, R., & Fast, N. J. (2021). Humans judge, algorithms nudge: The psychology of behavior tracking acceptance. *Organizational Behavior and Human Decision Processes, 164*, 11–26. https://doi.org/10.1016/j.obhdp.2021.01.001

Reuters. (2020, Mar 17). *Volkswagen says diesel scandal has cost it 31.3 billion euros*. Retrieved from https://www.reuters.com/article/us-volkswagen-results-diesel-idUSKBN2141JB

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rush, M. C., Thomas, J. C., & Lord, R. G. (1977). Implicit leadership theory: A potential threat to the internal validity of leader behavior questionnaires. *Organizational Behavior and Human Performance, 20*(1), 93–110. https://doi.org/10.1016/0030-5073(77)90046-0

Schyns, B., Kiefer, T., Kerschreiter, R., & Tymon, A. (2011). Teaching implicit leadership theories to develop leaders and leadership: How and why it can make a difference. *Academy of Management Learning & Education, 10*(3), 397–408. https://doi.org/10.5465/amle.2010.0015

Shamir, B. (2011). Leadership takes time: Some implications of (not) taking time seriously in leadership research. *The Leadership Quarterly, 22*(2), 307–315. https://doi.org/10.1016/j.leaqua.2011.02.006

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior, 86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Smith, I. H., Soderberg, A. T., Netchaeva, E., & Okhuysen, G. A. (2022). An examination of mind perception and moral reasoning in ethical decision-making: A mixed-methods approach. *Journal of Business Ethics. Advance online publication*. https://doi.org/10.1007/s10551-021-05022-9

Sullivan, Y. W., & Fosso Wamba, S. (2022). Moral judgments in the age of artificial intelligence. *Journal of Business Ethics. Advance online publication*. https://doi.org/10.1007/s10551-022-05053-w

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Miner, L., Wager, S., & Wright, M. (2018). GRF: Generalized random forests. Retrieved from https://CRAN.R-project.org/package=grf

van de Calseyde, P. P., Evans, A. M., & Demerouti, E. (2021). Leader decision speed as a signal of honesty. *The Leadership Quarterly, 32*(2), 1–11. https://doi.org/10.1016/j.leaqua.2020.101442

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14*(8), 383–388. https://doi.org/10.1016/j.tics.2010.05.006

Weber, L., & Mayer, K. J. (2011). Designing effective contracts: Exploring the influence of framing and expectations. *Academy of Management Review, 36*(1), 53–75. https://doi.org/10.5465/amr.2008.0270

Wesche, J. S., & Sonderegger, A. (2019). When computers take the lead: The automation of leadership. *Computers in Human Behavior, 101*, 197–209. https://doi.org/10.1016/j.chb.2019.07.027

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., et al. (2020). Transformers: State-of-the-art natural language processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics.

Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2020). Robots at work: People prefer-and forgive-service robots with perceived feelings. *Journal of Applied Psychology, 106*, 1557–1572. https://doi.org/10.1037/apl0000834

Yam, K. C., Goh, E.-Y., Fehr, R., Lee, R., Soh, H., & Gray, K. (2022). When your boss is a robot: Workers are more spiteful to robot supervisors that seem more human. *Journal of Experimental Social Psychology, 102*, 104360. https://doi.org/10.1016/j.jesp.2022.104360

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology, 85*, 103870. https://doi.org/10.1016/j.jesp.2019.103870