Organizational Behavior and Human Decision Processes 188 (2025) 104405

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

The transparency dilemma: How AI disclosure erodes trust

Oliver Schilke [a,b,*], Martin Reimann [c,d]

[a] Department of Management and Organizations, Eller College of Management, University of Arizona, United States
[b] Research Group in Leadership and Effective Organizations, EGADE Business School, Tecnológico de Monterrey, Mexico
[c] Department of Marketing, Eller College of Management, University of Arizona, United States
[d] Research Group in Consumer Behavior and Conscious Marketing, EGADE Business School, Tecnológico de Monterrey, Mexico

ARTICLE INFO

Editor: Mike Baer

Keywords:
Trust
Disclosure
Generative artificial intelligence
Legitimacy

ABSTRACT

As generative artificial intelligence (AI) has found its way into various work tasks, questions about whether its usage should be disclosed and the consequences of such disclosure have taken center stage in public and academic discourse on digital transparency. This article addresses this debate by asking: Does disclosing the usage of AI compromise trust in the user? We examine the impact of AI disclosure on trust across diverse tasks—from communications via analytics to artistry—and across individual actors such as supervisors, subordinates, professors, analysts, and creatives, as well as across organizational actors such as investment funds. Thirteen experiments consistently demonstrate that actors who disclose their AI usage are trusted less than those who do not. Drawing on micro-institutional theory, we argue that this reduction in trust can be explained by reduced perceptions of legitimacy, as shown across various experimental designs (Studies 6–8). Moreover, we demonstrate that this negative effect holds across different disclosure framings, above and beyond algorithm aversion, regardless of whether AI involvement is known, and regardless of whether disclosure is voluntary or mandatory, though it is comparatively weaker than the effect of third-party exposure (Studies 9–13). A within-paper meta-analysis suggests this trust penalty is attenuated but not eliminated among evaluators with favorable technology attitudes and perceptions of high AI accuracy. This article contributes to research on trust, AI, transparency, and legitimacy by showing that AI disclosure can harm social perceptions, emphasizing that transparency is not straightforwardly beneficial, and highlighting legitimacy's central role in trust formation.

1. Introduction

People are increasingly finding generative artificial intelligence (AI) to be a highly beneficial tool in facilitating their work. Nonetheless, there are widespread moral reservations about claiming authorship of AI-assisted work (Krausová & Moravec, 2022). As these technologies are becoming integral to organizations' daily operations (Kellogg et al., 2020), it is urgent to address the ethical considerations surrounding the usage of AI. The solution advocated in the popular press (e.g., Gay, 2024) and among ethics committees (e.g., Committee on Publication Ethics, 2023) seems straightforward: disclose the usage of AI.

However, in this article, we point to a hidden cost of AI disclosure—a loss in trust. We find that, despite being touted as a practice of ethical transparency, AI disclosure paradoxically erodes trust. This finding challenges prevailing assumptions and underscores the potential negative impact of transparency[1] on trust. We first argue that, ceteris paribus, AI disclosure will diminish trust in an actor across a variety of tasks. Second, we identify a mechanism through which AI disclosure affects trust, adding knowledge of how the AI disclosure–trust effect comes about. Third, we show that while AI disclosure jeopardizes trust, the exposure of undisclosed AI usage by a third party has an even more detrimental effect. Finally, in an effort to identify relevant boundary conditions, we point to individual differences among trustors that moderate the AI disclosure–trust effect.

The article makes several theoretical contributions. First, our investigation contributes to the burgeoning literature on the consequences of AI in organizations by pioneering the examination of social evaluations of AI usage. While recent research has focused on explaining the effectiveness of AI in making people more productive (Jia, Luo, Fang, & Liao, 2024; Noy & Zhang, 2023; Dell'Acqua et al., 2023), our

* Corresponding author at: The University of Arizona, McClelland Hall 405GG, 1130 E. Helen St, Tucson, AZ 85721-0108, United States.
E-mail address: oschilke@arizona.edu (O. Schilke).

[1] In this article, we employ the term *transparency* to refer to user transparency regarding users' application of AI rather than the conceptually distinct issue of AI system transparency.

https://doi.org/10.1016/j.obhdp.2025.104405
Received 20 June 2024; Received in revised form 10 March 2025; Accepted 13 March 2025
Available online 23 April 2025

article draws attention to the social dynamics of how the disclosure of AI usage affects how people are perceived. Interestingly, our meta-analytic findings did not show evidence of an attenuated AI-disclosure penalty among those who have used it themselves or are highly familiar with AI, which raises the possibility of its persistence even as AI continues to diffuse.

Second, our research contributes to the literature on transparency by calling into question the widely held belief that transparency uniformly yields favorable outcomes. The extant literature is dominated by a positive narrative, such that the notion that transparency enhances trust is often taken for granted (Schnackenberg & Tomlinson, 2016), with the assumption that openness invariably bolsters trust. However, the effect of transparency may be more contingent than commonly assumed (Sah et al., 2018) and may even reverse when the disclosed information compromises the discloser (Birkinshaw & Cable, 2017). Consistent with this view, our investigation demonstrates that transparency can backfire when one discloses AI usage. Our work shows remarkable robustness of such disclosure's trust-eroding effect across a wide variety of tasks and participants. Ironically, people who try to be trustworthy—by transparently disclosing AI usage—are trusted less.

Third, we contribute to the literature on trust by highlighting the critical role of legitimacy in the trust-erosion process. Drawing on micro-institutional theory (e.g., Bitektine & Haack, 2015; Zucker, 1977), we present a theoretical model in which trust judgments are heavily influenced by the extent to which a trustee's actions are deemed socially appropriate.

2. Trust and AI

It is now widely accepted that trust—the willingness to make oneself vulnerable to the actions of another party (Mayer et al., 1995)—is a key process that facilitates cooperation and coordination (Schilke et al., 2021). Trust is thus central to the effectiveness of various workplace relationships, including those between leaders and followers, interviewers and job applicants, negotiators, and teammates (Kramer, 1999). Consequently, trust has become the subject of a large stream of organizational research (Dirks & de Jong, 2022), with particular attention paid to how technological trends shape trust dynamics (e.g., Lumineau et al., 2023). As workplaces become increasingly digitalized, how advanced technology may transform relational processes has emerged as a central question in contemporary scholarship on trust (McKnight et al., 2011).

Most recently, the question of how humans form trust in AI agents has received considerable attention (see Glikson & Woolley, 2020 for a review of this growing research stream). In a widely adopted definition, the OECD (2024) describes AI as "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." Understanding how trust plays out in the context of AI is critically important considering that the scope of human interactions with AI is expanding drastically across various domains (Raisch & Krakowski, 2021). Recent studies have explored the conditions under which humans place either less or more trust in AI (Lockey & Gillespie, 2024), identifying several factors such as its anthropomorphism and reliability (Kaplan et al., 2023).

Most of this research places potential users in the role of the trustor, exploring how certain AI characteristics impact their trust in AI. While this perspective is important, there is a compelling argument to be made for considering AI users not only as trustors but also as trustees whose actions are judged by others. This dual role warrants further exploration of how evaluators perceive a trustee's engagement with AI. That is, with the increasing importance of AI, it is important to understand trust directed toward not only the technology itself but also those who use it.

2.1. AI disclosure and trust

In this article, we adopt the novel perspective described above, focusing on how a human's decision to disclose their usage of AI may impact the trust placed in them. The dilemma of whether to disclose AI usage has become a pressing concern for many, as evidenced by its widespread coverage in the popular press (e.g., Agarwal, 2023; Gay, 2024). Some view AI disclosure as a moral responsibility, as it enables people to clarify the role of AI in their work processes and give credit to the technology's contributions. On the other hand, people worry about how their disclosed usage of AI will be perceived, as reflected in recent polls: while most people believe that AI usage should be disclosed (Zetwerk, 2024), they hesitate to do so themselves (Fishbowl, 2023). Our investigation addresses the dilemma surrounding AI disclosure—understood as the act of informing audiences about the deployment of AI in one's work processes, products, or decisions. In particular, we investigate whether the concern that AI disclosure may lead to decreased trust is warranted.

As a starting point to motivate our argument, we note prior research into conflict-of-interest (COI) disclosure (Loewenstein et al., 2011; Sah, 2019), and we identify relevant parallels between COI and AI disclosure. Both COI and AI disclosures serve to provide transparency. Moreover, both types of disclosure acknowledge potentially significant and impactful factors external to the core decision-making entity: COI disclosures inform relevant parties about personal, financial, or institutional relationships that could influence decision making (Loewenstein et al., 2012), while AI disclosures aim to reveal the involvement of automated systems that could similarly influence processes or outcomes. Previous empirical research on COI disclosures has found that such transparency can erode trust in the actor making the disclosure (Sah & Feiler, 2020; Sah et al., 2018). Considering these findings, we propose that AI disclosure may trigger a similar reaction. We argue that AI disclosure serves as a warning to recipients that the discloser's work is not purely human-generated, which is likely to be viewed as illegitimate and consequently to diminish trust (as we will discuss in greater detail below). Thus, we hypothesize the following main effect:

Hypothesis 1: An actor disclosing (vs. not disclosing) the usage of AI for work tasks will be trusted less.

2.2. Legitimacy as a theoretical mechanism linking AI disclosure and trust

Next, we will draw on micro-institutional theory (e.g., Bitektine & Haack, 2015; Harmon, 2019b; Zucker, 1977; Zucker & Schilke, 2019) as a framework to unpack the AI disclosure–trust effect and discuss perceived legitimacy as a key mechanism. Micro-institutional theory is especially germane for our purposes as it helps explain the social dynamics in settings where nontraditional actions, like the disclosure of AI usage in professional settings, may conflict with taken-for-granted expectations (Zucker, 1987). The micro-institutional approach places an emphasis on the mental processes through which individuals make sense of social situations (Powell & Colyvas, 2008; Schilke, 2018; ir8511) and, in particular, the construction of legitimacy (Bitektine & Haack, 2015; Haack et al., 2021). Legitimacy refers to the perception that an entity's actions or decisions are desirable, proper, or appropriate in the given setting (Suchman, 1995). Past research has also referred to legitimacy using terms such as acceptability, taken-for-grantedness, appropriateness, expectation conformity, and congruence (Deephouse & Carter, 2005). Concerns about legitimacy tend to arise whenever individuals encounter practices that deviate from established norms or challenge their preconceived notions of appropriate behavior (Johnson et al., 2006).

Building on the notion that trust-building and trust-erosion processes can be seen as qualitatively distinct (Lewicki et al., 1998; Reimann et al., 2017b), we argue that legitimacy considerations are particularly impactful in situations of possible trust erosion, such as when AI usage is disclosed. In other words, we expect legitimacy to have asymmetric relevance in trust building versus trust erosion, with comparatively greater weight in the context of trust erosion. This view is consistent with recent micro-institutional research suggesting comparatively stronger effects of negative versus positive legitimacy cues (Schilke et al., forthcoming). A certain degree of consistency with normative expectations is a basic requirement that trustors expect trustees to display (Parsons, 1951), such that positive legitimacy judgments may have only marginal effects on trust building. In contrast, violations of the trustor's a priori expectations regarding appropriate behavior are a major driving force behind trust erosion (Elangovan et al., 2007; Lapidot et al., 2007). If a trustee acts in a way that raises doubts about their adherence to social norms and disrupts the natural course of things, as reflected by negative legitimacy judgments, this will trigger mental alarms (Harmon, 2019b) and drive trust erosion (Kramer, 1999). In other words, deviance from norms makes a trustee particularly susceptible to trust withdrawal.

One such widely held normative expectation relates to the locus of agency in the execution of tasks. In many work contexts, there is a strong assumption that people's decisions and outputs should be the result of human expertise, judgment, and reasoning (Palmeira & Spassova, 2015). This expectation is deeply rooted in cultural and legal norms that highly value humans' unique capabilities and insights in problem-solving and creative processes (Polanyi, 1966). When disclosed, AI involvement may be perceived as a deviation from these expectations, leading to a perception that the work practices are inappropriate because they diminish or even replace human agency (Martin & Waldman, 2023). This perceived deviation can undermine the legitimacy of work processes, as audiences may view the usage of AI as diminishing the valued human element traditionally associated with these processes. As such, we argue that disclosed AI usage will be perceived as inconsistent with socially established standards for task execution, resulting in diminished perceptions of legitimacy. In contrast, if AI usage remains undisclosed, no such perception shift occurs, allowing people to maintain a façade of conformity to accepted practices.

Importantly, it is the act of disclosing AI usage, rather than mere awareness of it, that can contribute to evaluators' focus towards scrutinizing the methodologies employed. As Harmon (2019b) demonstrates, transparency pledges meant to signal honesty and instill confidence may instead sow seeds of doubts. Openly disclosing practices to provide reassurance often draws heightened attention to them and raises questions about their appropriateness. Such disclosure, particularly when intended to preemptively dispel fears or doubts, can induce reactance, making evaluators more skeptical and resistant to the disclosed information (Brehm, 1966). In the terms of Toulmin's (1958) model of argument, making a claim (such as the disclosure of AI) may disrupt the taken-for-grantedness of a situation, prompting evaluators to examine and possibly challenge the claim's appropriateness—unlike situations where no such claim is made (Harmon, 2019a). Thus, while AI disclosure may aim to preempt misgivings, it can paradoxically invite greater scrutiny and skepticism about the legitimacy of the disclosing party's practices.

A lack of legitimacy, in turn, is an important contributor to trust erosion (Chen et al., 2022; Treviño et al., 2014). At its core, low legitimacy reflects perceptions of social inappropriateness and indicates divergence from social expectations. Because low legitimacy suggests that one fails to adhere to accepted norms and values, it seems unsafe for others to rely on the actions or directives of individuals who are perceived as illegitimate (Deephouse & Suchman, 2008). Thus, people are more likely to view such individuals in a negative light (Bitektine & Haack, 2015; Suddaby et al., 2017) and to withhold trust in their actions, decisions, and leadership. Illegitimate individuals' actions are perceived as unpredictable and not consistent with the roles they occupy, which creates a sense of insecurity that threatens others' trust in them. While trust may eventually have a feedback effect on legitimacy, we focus on the immediate effects of AI disclosure, justifying our emphasis on the legitimacy-to-trust direction. In sum, building on micro-institutional theory, we propose that perceived legitimacy serves as a critical link between AI disclosure and trust.

Hypothesis 2: Legitimacy mediates the negative effect of AI disclosure on trust, such that disclosing (vs. not disclosing) the usage of AI for work tasks reduces perceptions of legitimacy, which in turn erodes trust.

2.3. Conceptual variations of the independent variable in the model

While our core hypotheses (H1–2) build on micro-institutional theory to explain the legitimacy-based mechanisms linking AI disclosure to trust erosion, we also seek to examine the robustness of our findings across various conditions and introduce relevant variations to our independent variable. Specifically, we explore how several factors—the framing of the disclosure (H3), an autonomous-AI-agent baseline (H4), and the manner of AI usage revelation (H5)—may impact the observed effect. These hypotheses introduce conceptual variations that allow us to test whether the effect of AI disclosure on trust persists across different contexts.

Framing of AI disclosure. Thus far, we have treated AI disclosure as a uniform concept; however, it is important to acknowledge that such disclosure may be framed differently. According to micro-institutional theory, such variations in framing may affect how people construe the legitimacy of practices (Glaser et al., 2016; Harmon, 2019a), making it important to consider the language used in AI disclosure. In particular, the framing of AI disclosure can vary in terms of the specificity of the disclosure and the intended usage of AI that is being acknowledged (Ali et al., 2024). Disclosed usage of AI can be framed very generally (e.g., "This work task was prepared and processed by AI") or more specifically (e.g., noting that AI was used but that the human has revised the generated output or emphasizing that AI was used for proofreading only). The AI disclosure may also underscore the intent behind the AI usage, such as the desire to ensure high standards of written communication. It may also explicitly acknowledge that AI-generated content may contain errors, possibly to set realistic expectations. Finally, the AI disclosure could be framed with an emphasis on the purpose of the AI disclosure as an instrument for enhancing transparency. While different framing approaches may certainly have nuanced implications for trust formation, we posit that each of them will result in lower levels of trust (vs. no AI disclosure). We reason that this will occur because the mere acknowledgment of AI involvement, regardless of framing, can induce a sense of lacking legitimacy to evaluators, in turn leading to trust depreciation. Hence:

Hypothesis 3: An actor disclosing (vs. not disclosing) the usage of AI for work tasks will be trusted less, no matter whether the disclosure (a) is framed in general terms, or whether it includes a note (b) that a human has reviewed and revised the work, (c) that AI was used only for proofreading, (d) on the human's intent behind their AI usage, (e) that AI-generated content may contain errors, and (f) that the human is transparent about their AI usage.

AI disclosure vs. AI agent. One might suspect that the AI-disclosure effect is merely picking up on algorithm aversion, which manifests in more negative attitudes towards an algorithmic versus human agent (e.g., Dietvorst et al., 2018; Newman et al., 2020). However, we argue that AI disclosure goes further. Whereas algorithm aversion refers to a general skepticism toward algorithms due to possible errors, human disclosure of the involvement of AI introduces a legitimacy discount arising from role ambiguity, leading the human actor to be trusted even *less* than an autonomous AI agent performing the same task. This decrease in trust can be attributed to the uncertainty regarding the locus of agency when it is revealed that both a human and an AI were involved in the task, making it unclear who holds primary responsibility (Cañas, 2022). Drawing again on the micro-institutional perspective, clear roles provide a stable, taken-for-granted framework that ensures normative order and reduces ambiguity by embedding expectations and responsibilities within established institutional orders (e.g., Berger & Luckmann, 1966; Ocasio, 2023; Zucker, 1977). When these clear, predictable structures are disrupted—such as through the entanglement of human and AI agents—it challenges the consistency that comes from clearly defined roles and responsibilities. The lack of a singular, accountable agent leads to ambiguity, causing evaluators to perceive the situation as misaligned with normative institutional standards (Bovens, 2010). This blending of roles creates a perception of divergence from the well-defined norms that characterize both roles separately, thereby diminishing legitimacy and trust. In contrast, autonomous AI agents, when acting alone, occupy a well-defined institutional role as tools optimized for specific tasks (Logg et al., 2019). This role consistency reinforces the AI's legitimacy, as it meets the expectations of its designed function and institutional purpose. We hypothesize:

Hypothesis 4: A human actor disclosing the usage of AI for work tasks will be trusted less than an autonomous AI agent performing these work tasks.

AI disclosure vs. exposure. So far, we have focused on the user disclosing their own usage of AI, but what if this information comes from a different source? If the usage of AI is revealed by a third party, it can be seen as a serious violation of social norms, which can damage trust more severely than if expectations had been managed from the outset through voluntary self-disclosure. Micro-institutional scholars, such as Deephouse et al. (2017) and Zucker and Schilke (2019), argue that when actors manage communication proactively, such as by disclosing information voluntarily, they may maintain control over their legitimacy narratives. This reduces the risk of trust erosion that typically occurs when external parties disclose the information instead. Further support for this position comes from research on crisis communication, which finds that preemptive disclosure is less damaging than being exposed by a third party (Lee, 2016). For example, an organization that self-discloses a crisis before the media uncovers it leads onlookers to pay less critical attention to the media frenzy that may follow, resulting in less harm to the organization's reputation (Claeys et al., 2016). As such, we expect AI exposure to hurt trust more:

Hypothesis 5: An actor being exposed to having used AI for work tasks (vs. disclosing and vs. not disclosing having done so) will be trusted least.

3. Overview of studies

In the main text of this article, we report 13 experiments and a within-paper meta-analysis to examine the impact of disclosing AI usage on trust. In the experiments, participants took on the role of a trustor charged with assessing another actor, either an individual or a firm, that disclosed or did not disclose AI usage. One key goal across our set of studies was to examine the generalizability of the proposed effect of AI disclosure on trust across a variety of different tasks to establish that this effect is not narrowly confined to specific contexts but extends broadly.[2]

Study Structure and Methodology

Our selection of tasks was guided by recent reports about relevant organizational contexts in which AI usage is particularly common (McKinsey, 2023). We started by examining our key hypothesis—that actors disclosing (vs. not disclosing) the usage of AI for work tasks will be trusted less—with Study 1, which was conducted among university students, focusing on a professor's disclosure of using AI for grading her students. Studies 2–5 further tested our focal effect in other contexts, including job applications, advertisements, employee performance reviews, and work emails. We also tested our theorizing on why this effect occurs—that is, because of reduced legitimacy—through process-measurement, process-manipulation, and causal-chain research designs (Studies 6–8). Studies 9–13 addressed whether our effect depends on a particular framing of the AI disclosure; whether algorithm aversion is solely to blame for this effect; whether the effect persists when the trustor was aware that AI had been used before the disclosure; whether the effect depends on the disclosure being made voluntarily (vs. mandatorily); and whether trust deteriorates more when a third party (vs. the trustee) exposes the trustee's AI usage. Table 1 gives an overview of all studies.

We employed the experimental method throughout these studies. In addition to their recognized capacity to identify causal relationships and micro-level mechanisms (e.g., Levine et al., 2023; Podsakoff & Podsakoff, 2019), experiments have been recognized as a uniquely suitable method for research on disclosure (Sah & Feiler, 2020), legitimacy (Haack et al., 2021), and trust (Schilke et al., 2023). In particular, experiments allow us to isolate AI usage from AI disclosure, by holding constant the actual usage of the technology while cleanly manipulating its disclosure, and to assess the effect of this manipulation on trust.

3.1. Research transparency statement

The University of Arizona's Institutional Review Board provided approval for our research. We piloted all 13 experiments, which allowed us to conduct a priori power analyses to determine appropriate sample sizes. We then preregistered each of these studies on the Open Science Framework (OSF) prior to data collection (see Table 1 for individual links). These preregistrations include formal hypotheses, manipulations of the independent variables, measures of the dependent variables, planned sample size, exclusion criteria, and a plan for how the data would be analyzed.

After participants were recruited (via the behavioral lab for Study 1, in class for Study 8, or through the online CloudResearch Connect panel for all other studies), they provided their consent. Data collection was completed upon reaching the a priori-determined sample size and before commencing data analyses. Following the recommendation by Berinsky et al. (2014), no data were excluded from our experiments.[3]

All study materials—including instruments, data, syntax, outputs, and manipulation checks—are publicly available in the same OSF repositories where the preregistrations were posted—see Table 1. In addition, this article is accompanied by supplementary materials (*SM*). We adhered to the APA Style Journal Article Reporting Standards for quantitative studies (https://apastyle.apa.org/jars). We used Stata 17, SPSS 28, and the PROCESS package (Hayes, 2022) to analyze the data.

3.2. Study 1

The aim of Study 1 was to test H1—that an actor disclosing (vs. not disclosing) the usage of AI for work tasks will be trusted less—in a realistic environment. Disclosure of AI usage by a professor was manipulated and its impact on the students' trust in her was measured.

3.2.1. Participants and design

One hundred ninety-five undergraduate business students were recruited to participate in exchange for course credit (see the descriptive statistics of all sociodemographic variables in the *SM*). Participants were randomly assigned to one of three conditions in a one-factor between-subjects experimental design with three levels (AI disclosure, human-teaching-assistant disclosure, no-disclosure control). We added a

[2] To examine the nature of the tasks used herein, we conducted a study in which participants categorized the tasks along eight different dimensions such as task objectivity versus subjectivity. Results showed substantial variation in how participants rated the tasks across these dimensions. For example, composing emails is viewed as rather objective, whereas conducting market research is viewed as rather subjective. Details are reported in Study SM-1 in the *SM*. Results underscore the diversity of our investigation's tasks in terms of perceived cognitive, conative, and affective demands, which in turn reinforces the generalizability of our findings across various contexts.

[3] In exploratory post-hoc analyses reported in the individual experiments' log files, we dropped participants who provided incorrect responses to any of the three attention screeners, and results were substantially similar.

Table 1
Study Overview.

Study	Preregistration and repository	Hypothesis addressed	Type of effect	Experimental manipulation(s)	Dependent variable(s)	Sample	Key finding(s)
1	https://osf.io/w8acv	H1	Main effect	AI disclosure, human-teaching-assistant disclosure, no-disclosure control	Trust in teacher	195 students	Students trusted the professor less when disclosing AI usage for grading (vs. disclosing usage of a human graduate teaching assistant or making no disclosure).
2	https://osf.io/svp9n	H1	Main effect	AI disclosure, human-career-coach disclosure, no-disclosure control	Trust Likelihood of hiring job applicant	85 supervisors with hiring experience	Results of this study generalize to trustors in higher-power position (vs. lower-power position in Study 1), to the context of AI usage for writing a job application, and to a behavioral intention measure of trust.
3	https://osf.io/wvspg	H1	Main effect	AI disclosure, human-financial-analyst disclosure, no-disclosure control	Trust in firm Willingness to invest Amount to invest (US$)	345 investors	Results of this study generalize to firm trustees (vs. individual trustees in Studies 1 and 2) and to the context of AI usage for creating an advertisement.
4	https://osf.io/cpfxj	H1	Main effect	AI disclosure, I-disclosure, no-disclosure control	Trust	90 legal analysts	Results of this study generalize to disclosing one's own effort (vs. disclosing that work was outsourced to other humans in Studies 1–3) and to the context of AI usage for writing an annual performance review.
5	https://osf.io/jfe9a	H1	Main effect	AI disclosure, no-disclosure control	Trust	597 panel members	Results of this study generalize to the everyday task of composing an email to a coworker.
6	https://osf.io/yjmtp	H1, H2	Indirect effect in a measure-ment-of-mediation design	2 (AI disclosure, no-disclosure control) × 2 (valence of decision context: termination, employment)	Trust Legitimacy	427 panel members	Legitimacy mediates the negative effect of AI disclosure on trust in a measurement-of-process design. Results of this study generalize to the context of AI usage for writing letters of termination and of employment. The AI disclosure–trust effect is robust across negative and positive settings.
7	https://osf.io/94c7j	H1, H2	Mediation-by-moderation effect	2 (AI disclosure, no-disclosure control) × 2 (collective validity: prime, control)	Trust	426 panel members	Legitimacy mediates the negative effect of AI disclosure on trust in a manipulation-of-process design. Results of this study generalize to the context of AI usage for writing a bio sketch.
8	https://osf.io/4w2zj	H2	Main effect in a causal- chain mediation design	High legitimacy, low legitimacy	Amount to invest (US$) Trust	93 students	Legitimacy has a causal effect on trust. Results of this study generalize to a behavioral measure of trust involving financial investment.
9	https://osf.io/kxj2d	H1, H3	Main effect	Six different AI disclosure framings, no-disclosure control	Trust	518 panel members	An actor disclosing (vs. not disclosing) the usage of AI for work tasks will be trusted less, regardless of whether the disclosure (a) is framed in general terms or, instead, includes a note (b) that a human has reviewed and revised the work, (c) that AI was used only for proofreading, (d) that the human's intent in using AI was to enhance writing quality, (e) that AI-generated content may contain errors, or (f) emphasizing the importance of transparency about AI usage.
10	https://osf.io/je8sw	H1, H4	Main effect	Autonomous-AI agent, Human-actor-with-AI disclosure, human-actor-no-disclosure control	Trust in message	753 panel members	A human actor disclosing the usage of AI for work tasks will be trusted less than an autonomous AI agent performing these work tasks. Results of this study generalize to the context of AI usage for generating health and safety guidelines.
11	https://osf.io/4a8ds	H1	Main effect	2 (AI disclosure, no-disclosure control) × 2 (evaluator's knowledge of AI usage: present, control)	Cognition-based trust Affect-based trust	1,048 panel members	A human actor disclosing the usage of AI for work tasks will be trusted less, regardless of whether the AI usage is known or not known by the evaluator prior to disclosure. Results of this study generalize to the context of AI usage for market research.
12	https://osf.io/mxfc2	H1	Main effect	2 (AI disclosure, no-disclosure control) × 2 (disclosure regime:	Consumer trust Willingness to rehire	348 panel members	An actor disclosing (vs. not disclosing) the usage of AI for work tasks will be trusted less, regardless of whether the disclosure is made voluntarily or

(*continued on next page*)

Table 1 (*continued*)

Study	Preregistration and repository	Hypothesis addressed	Type of effect	Experimental manipulation(s)	Dependent variable(s)	Sample	Key finding(s)
				voluntary, mandatory)			mandated by regulation. Results of this study generalize to the context of AI usage for graphic design.
13	https://osf.io/ve8r5	H1, H5	Main effect	AI disclosure, AI exposure, no-disclosure control	Trust	195 panel members	An actor exposed for using AI for work tasks will be trusted less than an actor disclosing such AI usage. Results of this study generalize to the context of AI usage for generating tax return advice.
Within-paper meta-analysis	https://osf.io/kng62 (repository only)	H1	Main effect, exploratory moderation effects	n/a	Trust	4,093 individuals	The trustor's attitude toward technology and perception of AI accuracy moderate the effect of AI disclosure on trust.

Notes. Throughout Studies 2–7 and 9–13, we recruited all online panelists from the United States using the CloudResearch Connect platform in exchange for monetary compensation.

human-assistant condition to the experiment to control for the possibility of our focal effect being simply due to the act of delegating work. For succinctness, refer to Table 1 for information on participants and design for all other experiments.

3.2.2. Experimental procedure

Participants were invited to the behavioral lab, where they were exposed to the online learning platform of their university, called Desire2Learn. Participants read that, while trying to get a sense of their course load at the beginning of the spring semester, they came across a welcome message from one of their professors, Elena Richardson, teaching the course Managing Groups and Teams. In this and all other experiments, participants responded to true-or-false comprehension questions to probe whether they had understood the scenario. Participants were then randomly assigned to one of three conditions and advanced to the next screen of the online learning platform with the professor's message, which either stated that all assignments for this course would be automatically graded by generative AI (AI-disclosure condition) or by a human graduate teaching assistant (human-teaching-assistant-disclosure condition) or made no such statement (no-disclosure condition). The effectiveness of the manipulation was tested with an independent sample ($n = 48$) to avoid potential demand effects (Podsakoff & Podsakoff, 2019). The manipulation check revealed that our experimental manipulation was successful. Participants assigned to the AI-disclosure condition exhibited markedly higher mean values across our three manipulation check items (which assessed clarity, transparency, and awareness of the professor's AI disclosure on seven-point answer scales, $\alpha = 0.99$) compared to those assigned to either the human-teaching-assistant-disclosure condition ($p < 0.001$) or the no-disclosure condition ($p < 0.001$). For succinctness, we will not mention the manipulation checks for the remaining experiments but report them in the *SM*. After reading the welcome message, participants were asked about their first impression of the professor based on the eight-item trust-in-teacher scale (e.g., "I can trust the way this teacher uses his or her power and authority") (Gregory & Ripski, 2008). For additional succinctness, we report the full lists of items of all our dependent variables, their answer scales, and their scale reliabilities in the *SM*. Finally, we collected additional information at the end of each of our studies; for succinctness, we will only mention these measures here and not for the remaining studies. In particular, participants briefly elaborated on their trust assessment in an open-ended text entry field; reported their sex assigned at birth, age, and race for demographic purposes (see the descriptive statistics in the *SM*); and responded to an attention check about the study's topic. Participants were also asked about their relationship with advanced technology (i.e., familiarity with AI, attitude towards new technological advancements, prior AI usage at work, and belief in the reliability of AI), which we will address in our within-paper meta-analysis.

3.2.3. Results

Analysis of variance revealed an effect of disclosure on trust, $F(2, 192) = 15.92$, $p < 0.001$, $\eta^2 = 0.14$. In support of H1, pairwise-comparison t tests revealed that the undergraduate students trusted their professor less when she disclosed employing AI for the grading of their class assignments ($M = 2.48$, $SD = 0.56$) than when she disclosed employing a human graduate teaching assistant to do so ($M = 2.87$, $SD = 0.46$, $t(130) = 4.44$, $p < 0.001$, Cohen's $d = 0.77$) or made no such disclosure ($M = 2.96$, $SD = 0.54$; $t(127) = 4.99$, $p < 0.001$, $d = 0.88$). The difference between the human-teaching-assistant-disclosure and no-disclosure conditions was small and not statistically significant by conventional standards ($t(127) = 0.99$, $p = 0.33$, $d = 0.17$).

3.3. Study 2

The aim of Study 2 was to determine whether our findings generalize by testing H1 in the context of hiring a job applicant. Disclosure of AI usage by a job applicant was manipulated and its impact on the hiring manager's trust in the applicant was measured. This design also allowed us to flip the structural power of the trustor (i.e., the participant) and the trustee (Schilke et al., 2015). Whereas the student participants in Study 1 were in a less powerful position than their professor, in Study 2 the participants assumed the role of a hiring manager with hiring power over the applicant.

3.3.1. Experimental procedure

Building on procedures previously established by Kim et al. (2004), we designed a task in which participants assumed the role of a manager in charge of hiring a senior-level tax accountant at the firm Michael Blankstein Tax Accounting Service. Participants were informed that they would make their hiring decision based on letters of motivation. Participants next read a letter of motivation from a job applicant named Ballou Mayers and were then randomly assigned to one of three conditions. Depending on the condition, participants either read a sentence indicating that the letter was prepared by generative AI (AI-disclosure condition) or by a human career coach and human resource expert named M. Zanger (human-career-coach-disclosure condition) or were shown a spinning wheel for several seconds indicating that they had to wait (no-disclosure-control condition). After reading the letter, participants indicated the degree of their trust in the job applicant on a four-item trust scale adapted from Mayer and Davis (1999), with an example item being "I would be willing to let Ballou have complete control over my future in this company."[4] Participants also rated how likely they would be to hire the applicant, on a scale ranging from 1 = *definitely not* to 7 = *definitely*, serving as a behavioral-intention measure of trust (Kim et al., 2004). In addition to the other demographics collected in Study 1, Study 2 and the studies that follow also collected information on education, income, and work experience (see *SM*).

3.3.2. Results

Analysis of variance revealed an effect of disclosure on trust, $F(2, 82) = 14.39$, $p < 0.001$, $\eta^2 = 0.26$. In further support of H1, pairwise-comparison *t* tests revealed that participants trusted the job applicant less when the motivation letter disclosed employing generative AI ($M = 2.91$, $SD = 1.14$) compared to when it disclosed employing a career coach ($M = 4.32$, $SD = 1.39$, $t(54) = 4.15$, $p < 0.001$, $d = 1.11$) or made no such disclosure ($M = 4.40$, $SD = 0.95$; $t(55) = 5.35$, $p < 0.001$, $d = 1.42$). The difference between the human-career-coach-disclosure and no-disclosure conditions was small and not statistically significant by conventional standards ($t(55) = 0.24$, $p = 0.81$, $d = 0.06$). A highly similar pattern was also found for participants' likelihood of hiring the applicant ($F(2, 82) = 13.60$, $p < 0.001$, $\eta^2 = 0.25$).

3.4. Study 3

The aim of Study 3 was to further generalize our findings by testing H1 in yet another context: trust in an investment fund's advertising for one of its financial products. Disclosure of AI usage by an investment fund was manipulated and its impact on the prospective investor's trust in the fund was measured. We recruited subjects who had one or more of the following holdings: cash in checking or savings accounts, certificates of deposit, stocks, mutual funds or electronically traded funds, 401ks or IRAs, bonds, and/or real estate. The design of this study also allowed us to determine whether H1 is supported when people are asked about their trust in a business entity rather than in a person (as in Studies 1 and 2).

3.4.1. Experimental procedure

Following Koehler and Mercer (2009), we asked participants to imagine they were browsing a popular business publication and came across the advertisement of a mid-sized investment company called Allen Funds, which interested them because they were trying to save some money for future use. Their task was to study the ad and later make an investment decision. Depending on the condition, they either read that the ad was outsourced and prepared by generative AI (AI-disclosure condition) or by a human financial analyst, Thomas Fischer, of another firm called Financial Services LLC (human-financial-analyst-disclosure condition) or were shown a spinning wheel (no-disclosure condition). Next, participants responded to a one-item trust measure stating "Allen Funds is an investment company that deserves investors' trust" on a 7-point Likert scale, ranging from 1 = *strongly disagree* to 7 = *strongly agree* (Koehler & Mercer, 2009). On the following screen, participants read: "Allen Funds plans to introduce a new growth fund that would have the same type of quality management team that you have come to expect from our funds." They were then asked how willing they were to invest a portion of a $10,000 gift in this new fund (1 = *definitely not willing*; 7 = *definitely willing*). Additionally, they were asked to indicate what percentage of the $10,000 they would be willing to invest in this fund on a slider measure from 0 % to 100 % (Koehler & Mercer, 2009). Toward the end of the study, participants also responded to several items intended to gauge their financial literacy. The items asked whether they had personal investment experience (63.48 % said yes; $M = 5.95$ years of experience, $SD = 8.45$), planned to invest in the near future (1 = *extremely unlikely*; 7 = *extremely likely*; $M = 5.29$, $SD = 1.80$), and made professional investment decisions for others (97.10 % said no; $M = 0.07$ years of experience, $SD = 0.52$), as well as how many hours per week they studied financial information ($M = 1.60$ h, $SD = 2.40$).

3.4.2. Results

Analysis of variance revealed an effect of disclosure on trust, $F(2, 342) = 17.43$, $p < 0.001$, $\eta^2 = 0.09$. In further support of H1, pairwise-comparison *t* tests revealed that investors trusted the investment fund less when the ad disclosed outsourcing its preparation to generative AI ($M = 4.18$, $SD = 1.45$) compared to human-financial-analyst disclosure ($M = 4.90$, $SD = 1.05$, $t(237) = 4.43$, $p < 0.001$, $d = 0.57$) and no disclosure ($M = 5.08$, $SD = 1.14$; $t(223) = 5.14$, $p < 0.001$, $d = 0.69$). The difference between the human-financial-analyst-disclosure and no-disclosure conditions was small and not statistically significant by conventional standards ($t(224) = 1.21$, $p = 0.23$, $d = 0.16$). Highly similar patterns were also found for the willingness to invest ($F(2, 342) = 12.65$, $p < 0.001$, $\eta^2 = 0.07$) and the amount invested ($F(2, 342) = 11.48$, $p < 0.001$, $\eta^2 = 0.06$).

3.5. Study 4

The aim of Study 4 was to test H1 among legal professionals evaluating a supervisor after receiving an annual performance review. In this study, we specifically recruited subjects with experience working in law firms. The supervisor's disclosure of AI usage was manipulated and its impact on employees' trust in the supervisor was measured. Compared to the previous studies, the design of this study differed in two ways: (1) by disclosing AI usage at the beginning of the written communication (vs. at the bottom or on the next screen) and (2) by introducing an I-disclosure condition, as opposed to not explicitly mentioning that the author had personally prepared the written communication (i.e., the control condition in Study 1 made no mention of the author's identity, and the control condition in Studies 2 and 3 showed a spinning wheel) or had outsourced the written communication to another human (i.e., teaching assistant, career coach, or analyst).

3.5.1. Experimental procedure

We designed a task in which participants imagined they were employees at a small law firm called Jones Skidd Webber Law Office. Their work included conducting legal research, drafting documents, and providing advice on legal matters. Participants were also told that they reported to a partner at the law firm, Thomas Skidd, JD, who would send them their annual performance evaluation on the day of the study. Depending on the condition, participants read their performance review, which either stated that it had been prepared by generative AI (AI-disclosure condition), explicitly stated that it had been prepared by their supervisor (I-disclosure condition), or included no such statement (control condition). After reading the performance review, participants were asked to indicate the degree of trust they placed in Thomas (adapted from Mayer & Davis, 1999).

3.5.2. Results

Analysis of variance revealed an effect of disclosure on trust, $F(2, 87) = 10.75$, $p < 0.001$, $\eta^2 = 0.20$. In further support of H1, pairwise-comparison *t* tests revealed that participants trusted the supervisor less when their annual performance review stated that it had been prepared by generative AI ($M = 3.80$, $SD = 1.59$) than when it explicitly stated that it had been prepared by their supervisor ($M = 5.11$, $SD = 0.83$, $t(57) = 3.90$, $p < 0.001$, $d = 1.02$) or made no such statement ($M = 5.00$, $SD = 1.09$; $t(60) = 3.48$, $p < 0.001$, $d = 0.88$). The difference between the I-disclosure and no-disclosure conditions was small and not statistically significant by conventional standards ($t(57) = 0.42$, $p = 0.68$, $d = 0.11$).

[4] Note that we adapted one item from the original four-item trust scale by Mayer and Davis (1999). Specifically, we replaced the item "I really wish I had a good way to keep an eye on Ballou" with "I trust Ballou." To ensure this modification did not affect the validity of our results, we conducted a replication of Study 2 with the original four items, which reproduced Study 2's results (see Study SM-2 in the *SM*).

3.6. Study 5

The aim of Study 5 was to test H1 in the context of an everyday, mundane task: sending scheduling emails to coworkers.

3.6.1. Experimental procedure

We designed a task in which participants assumed the role of a member of a creative team at a design firm called Design Space that specializes in innovative retail spaces. Their role was to work closely with other designers, brand managers, and technical staff to ensure the proposed store design in a current project is both attractive and functional. One of their coworkers, Karen Sinclair, was their key collaborator on this project. Participants were told to expect an email from Karen in their inbox. After waiting for their inbox to open, participants read Karen's mail. Depending on the condition, participants either read that Karen had used Superhuman, an AI tool, to assist in managing her calendar and sending scheduling emails (AI-disclosure condition) or were shown a spinning wheel (no-disclosure-control condition). Next, participants were asked to indicate the degree of trust they placed in Karen (adapted from Mayer & Davis, 1999).

3.6.2. Results

In further support of H1, an independent-samples *t* test revealed that participants trusted Karen less when she disclosed using generative AI for emailing ($M = 4.13$, $SD = 1.07$) compared to no such disclosure ($M = 4.49$, $SD = 0.91$; $t(595) = 4.43$, $p < 0.001$, $d = 0.36$).

3.7. Study 6

The aim of Study 6 was to test H2, which states that legitimacy mediates the negative effect of disclosure on trust, such that disclosing (vs. not disclosing) the usage of AI for work tasks reduces perceptions of legitimacy, which in turn erodes trust. Disclosure of AI usage by a warehouse worker's managing director was manipulated and its impact on the worker's trust in him was measured. The 2 × 2 design of this study also allowed us to determine whether H1 holds for both positively and negatively valenced decision contexts. While our previous studies' content was generally positive (e.g., a professor's welcome message, a job applicant's motivation letter, or a firm's advertisement), this study included both a highly positive and a highly negative type of correspondence—respectively, a letter of employment and a letter of termination.

3.7.1. Experimental procedure

We designed a task in which participants supposed they were one of the workers in a warehouse company called Dock Logistics Storage & Fulfillment, Inc. Participants were told that, as a warehouse worker, they were part of a team and played a crucial role in the smooth warehouse operations. After reading the job description, participants read a letter recently sent from the managing director, Alexander Vanderberg, to Zac Mayers. Depending on the valence condition, Zac was either a job applicant for a position as a warehouse worker like the participant or a current co-worker of the participant. In the positively valenced condition, the letter from Alexander to Zac was a letter of employment, whereas in the negatively valenced condition it was a letter of termination. Depending on the disclosure condition, participants either read that the letter was prepared by generative AI (AI-disclosure condition) or were shown a spinning wheel (no-disclosure-control condition). After reading the letter, participants responded to a twelve-item measure to indicate their perceptions of the managing director's legitimacy, with an example item being "The general public approves of the managing director's procedures" (Elsbach, 1994). Participants then indicated the degree to which they trusted the managing director (adapted from Mayer & Davis, 1999).

3.7.2. Results

Analysis of variance revealed an effect of disclosure on trust, $F(1, 423) = 74.91$, $p < 0.001$, $\eta^2 = 0.15$; an effect of valence on trust, $F(1, 423) = 52.34$, $p < 0.001$, $\eta^2 = 0.11$; and an interaction effect of disclosure and valence on trust, $F(1, 423) = 3.69$, $p = 0.06$, $\eta^2 = 0.01$. In further support of H1, an independent-samples *t* test revealed that participants trusted the managing director less when he disclosed using generative AI to prepare the letter ($M = 2.60$, $SD = 1.00$) than when he made no such disclosure ($M = 3.28$, $SD = 0.74$; $t(425) = 7.87$, $p < 0.001$, $d = 0.76$). Interestingly, the main effect of disclosure on trust was somewhat stronger in the negatively valenced condition (i.e., when terminating; $d = 0.92$) than in the positively valenced condition (i.e., when hiring; $d = 0.76$). However, our focal effect is robust across positively and negatively valenced decision contexts. In support of H2, mediation analysis using the PROCESS macro (model 4) with 5,000 bootstrapped samples (Hayes, 2022) revealed that perceptions of the managing director's legitimacy mediated the effect of disclosure on trust ($ab = -0.56$, $SE = 0.06$, 95 % CI: [−0.68, −0.43]). AI disclosure reduced perceptions of legitimacy ($a = -0.60$, $SE = 0.07$, 95 % CI: [−0.74, −0.46]), and perceived legitimacy in turn was positively related to trust ($b = 0.93$, $SE = 0.04$, 95 % CI: [0.85, 1.00]). This mediation effect holds both in the negatively valenced condition ($ab = -0.71$, $SE = 0.10$, 95 % CI: [−0.90, −0.52]), and in the positively valenced condition ($ab = -0.38$, $SE = 0.07$, 95 % CI: [−0.54, −0.25])

3.8. Study 7

The aim of Study 7 was to provide further support for H2. To do so, we used a 2 × 2 mediation-by-moderation experimental design (Spencer et al., 2005) in which both disclosure and collective validity were manipulated and their impact on trust was measured. A mediation-by-moderation approach suggests that a variable that affects the proposed psychological process (here, legitimacy) should interact with the independent variable. The notable strength of this approach is that it can contribute causal evidence for mediation (Podsakoff & Podsakoff, 2019).

A widely accepted driver of perceived legitimacy is collective validity (Haack et al., 2021; Suddaby et al., 2017), understood as the extent to which a social collective considers an individual's actions to be appropriate. When validity cues have a positive valence, such that an individual's practices are known to be viewed favorably by relevant authorities or peers, they will increase legitimacy perceptions (Zelditch & Walker, 1984). If legitimacy is a relevant mechanism explaining the link between AI disclosure on trust, then collective validity will moderate the negative effect of AI disclosure on trust, such that the effect is attenuated when collective validity is higher (vs. lower). Our manipulation of collective validity follows Miller and Morrison (2009) in reasoning that if the majority of people engage in a behavior, then others will assume it must be valid (also see Schilke & Rossman, 2018). The design of this study also allowed us to generalize our account to the context of technology startup founders, offering a conservative test for H1 given the prevalence of advanced technology usage in this field.

3.8.1. Experimental procedure

Depending on the collective-validity condition, participants either were or were not shown an excerpt from an actual article, published on the website of Boston Consulting Group's Henderson Institute (Candelon et al., 2023), discussing how people can create value with generative AI. The article notes that generative AI can improve people's performance by 40 % on tasks involving ideation and content creation. A chart was shown in support of this finding. Based on the work of Holtz (2015), all participants were then subjected to a task in which they assumed the role of an employees of a small technology startup company called Sigma, which offers search engine optimization services to other firms. Next, participants read the biographical profile of Chris J. Smith, Sigma's founder and their boss; then, depending on the disclosure condition, either the profile disclosed the usage of generative AI for its preparation (AI-disclosure condition) or participants were shown a spinning wheel (no-disclosure-control condition). After reading the biographical profile, participants responded to a four-item measure to indicate their degree of trust they would place in the startup founder (Holtz, 2015).

3.8.2. Results

Analysis of variance revealed a main effect of disclosure on trust, $F(1, 422) = 89.88$, $p < 0.001$, $\eta^2 = 0.18$; no main effect of collective validity on trust, $F(1, 422) = 0.42$, $p = 0.52$, $\eta^2 = 0.00$; and an interaction effect of disclosure and collective validity on trust, $F(1, 422) = 3.93$, $p = 0.048$, $\eta^2 = 0.01$. In further support of H1, an independent-samples *t* test revealed that participants trusted the startup founder less when his biographical profile disclosed the usage of generative AI ($M = 4.55$, $SD = 1.34$) than when no such disclosure was made ($M = 5.63$, $SD = 0.94$; $t(424) = 9.61$, $p < 0.001$, $d = 0.93$). In support of H2, the main effect of disclosure on trust was substantially stronger in the collective-validity-control condition ($d = 1.12$) than in the collective-validity-prime condition ($d = 0.72$), consistent with the hypothesized interaction effect. Fig. 1 illustrates this result. This moderation effect supports the notion that our focal effect is explained by legitimacy.

3.9. Study 8

The aim of Study 8 was twofold. First, following the recommendation by Spencer et al. (2005), we complemented Study 6′s measurement-of-process design and Study 7′s manipulation-of-process-design with an experimental-causal-chain approach, testing for the causal link from legitimacy to trust, as implied by H2. Second, we employed a behavioral (rather than perceptual or intentional) measure of trust to probe whether results generalize to settings with real monetary consequences (Levine et al., 2023).

3.9.1. Experimental procedure

In the classroom, students were randomly assigned to either a high- or low-legitimacy condition. They received a paper-and-pencil study packet containing instructions, the experimental manipulation, all measures, and an envelope with $5 in cash. They were instructed to take the money and place it in their pocket or wallet. On the next page, it was announced that they would be given an opportunity to invest the money in an actual exchange-traded fund (ETF). Before receiving further details about the ETF, students read a message from a stock market expert, who is a professor at their college. In the high-legitimacy condition, the professor’s message included a statement that “(…) investment vehicles that align with regulatory standards, industry norms, and institutional expectations tend to be viewed as highly legitimate. This ETF meets these criteria, including transparency of its investment strategy and consistencies in its historical performance.” By contrast, in the low-legitimacy condition, the professor’s message emphasized that “(…) investment vehicles that fail to align with regulatory standards, industry norms, and institutional expectations tend to raise serious legitimacy concerns. This ETF exhibits those warning signs, including a lack of transparency of its investment strategy and inconsistencies in its historical performance.” Across the two conditions, the messages were nearly identical in length and wording, differing only in their stance on the ETF’s legitimacy. Participants then read information from the actual prospectus of the fund indicating it invests in growth stocks of 1,000 large U.S. companies. Participants then made their investment decision, knowing that the experimenter would invest their chosen amount by the end of day and that they would receive their principle back on the target date one month out, plus or minus any gains or losses. Participants then responded to “This ETF is an investment fund that deserves investors’ trust” (0 = *strongly disagree*; 5 = *strongly agree*) and “How much are you investing in this ETF today?” *($0 − $1 − $2 − $3 − $4 − $5)*. Lastly, participants placed their investment amount, if any, into the envelope. The experimenter then purchased the ETF in the exact amount participants had placed in their envelopes ($346 in total).

3.9.2. Results

In further support of H2, an independent-samples *t* test revealed that participants invested less into the ETF when it was deemed illegitimate ($M = \$3.39$, $SD = 1.86$) compared to when it was deemed legitimate ($M = \$4.09$, $SD = 1.20$; $t(91) = 2.14$, $p = 0.03$, $d = 0.44$). A highly similar pattern emerged for the perceptual trust item ($t(91) = 4.76$, $p < 0.001$, $d = 0.99$).

3.10. Study 9

The aim of Study 9 was to test H3 that an actor disclosing (vs. not disclosing) the usage of AI for work tasks will be trusted less, regardless of whether the disclosure (a) is framed in general terms (as in the prior studies) or includes a note (b) that a human has reviewed and revised the work, (c) that AI was used only for proofreading, (d) regarding the human’s intent behind their AI usage, (e) that AI-generated content may contain errors, or (f) highlighting the human’s transparency about their

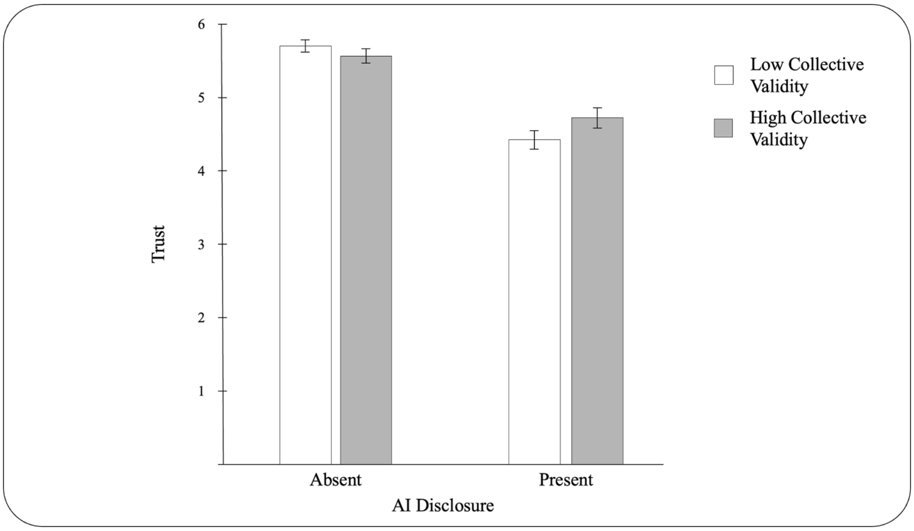

Fig. 1. Study 7: The Negative Effect of AI Disclosure on Trust Is Moderated by Collective Validity, Such That the Effect Will Be Attenuated When Collective Validity Is Higher (Vs. Lower).

AI usage.

3.10.1. Experimental procedure

The design of this study was identical to that of Study 2 except that we included five additional ways AI usage disclosure can be framed by a human user and omitted the career-coach condition. Participants read a letter from a job applicant, which, depending on their randomly assigned condition, disclosed the usage of AI in one of six different ways or made no such disclosure. The *SM* list the exact wording of each disclosure condition. Participants then indicated their level of trust in the job applicant (adapted from Mayer & Davis, 1999) and rated how likely they would be to hire the job applicant.

3.10.2. Results

Analysis of variance revealed an effect of disclosure on trust, $F(6, 511) = 15.79$, $p < 0.001$, $\eta^2 = 0.16$. In support of H3, pairwise-comparison t tests revealed that participants trusted the job applicant less when they disclosed employing generative AI in all six disclosure conditions compared to control (all $ps < 0.02$; see *SM* for exact test statistics). A highly similar pattern was found for the likelihood of hiring the job applicant ($F(6, 511) = 21.04$, $p < 0.001$, $\eta^2 = 0.20$).

3.11. Study 10

The aim of Study 10 was to test H4 that a human actor disclosing the usage of AI for work tasks will be trusted less than an autonomous AI agent performing these work tasks. We used a controlled experimental design in which the author of an email was either an autonomous AI agent or a human actor who disclosed or did not disclose AI usage for writing the email. We measured the email recipient's trust in the author of the email.

3.11.1. Experimental procedure

We designed a task in which participants assumed the role of a production line worker in a food and beverage processing plant called Harvest Gold Foods & Beverages Ltd. Participants read about their duties and were then told to expect an email with an update on the plant's health and safety guidelines. Participants then received and read the email, which, depending on the condition, was authored either by a bot (autonomous-AI-agent condition) or by the company's employee health and safety advisor, Ethan Marshall, who either disclosed (human-actor-with-AI-disclosure condition) or did not disclose the usage of AI for writing the email (human-actor-no-disclosure control). Participants were then asked about the degree of trust they placed in its author's message, based on a five-item scale, with an example item being "believable" (Flanagin & Metzger, 2000).

3.11.2. Results

Analysis of variance revealed an effect of disclosure on trust, $F(2, 750) = 46.07$, $p < 0.001$, $\eta^2 = 0.11$. In support of H4 and H1, pairwise-comparison t tests revealed that participants trusted the email message least when it was authored by a human actor who disclosed the usage of AI ($M = 4.94$, $SD = 1.15$) compared to when it was authored by an autonomous AI agent ($M = 5.61$, $SD = 0.92$, $t(506) = 7.27$, $p < 0.001$, $d = 0.64$) or a human actor who made no disclosure ($M = 5.72$, $SD = 0.88$; $t(496) = 8.56$, $p < 0.001$, $d = 0.77$). The difference between the autonomous-AI-agent and human-actor-no-disclosure-control conditions was small and not statistically significant by conventional standards ($t(498) = 1.41$, $p = 0.16$, $d = 0.13$).

3.12. Study 11

The aim of Study 11 was to test the robustness of H1 regardless of whether the AI usage is known or not known by the evaluator prior to disclosure. The evaluator's knowledge of AI usage and the trustee's disclosure of AI usage were manipulated, and their effects on trust were measured. This 2 × 2 design allowed us to disentangle the effect of disclosure from the effect of AI usage itself, providing a more nuanced understanding of how disclosure influences trust in the context of AI. As an additional robustness check, we utilized an alternate operationalization of trust that McEvily and Tortoriello's (2011) review revealed has been extraordinarily popular in the literature—McAllister's (1995) affect- and cognition-based trust measures.[5]

3.12.1. Experimental procedure

We designed a task in which participants assumed the role of head of marketing of Lumiere Studio, an entertainment company specializing in film and television production. Participants read about their duties, which included working with one of their team members, Alex Harper, a market research analyst. All participants were told that they were about to evaluate and act on the monthly competitor analysis report prepared by Alex. Depending on the experimental condition, participants either were also told that Alex employs AI tools to summarize the strategies, performance, and audience engagement of Lumiere Studio's competitors (evaluator's-knowledge-of-AI-use-present condition) or were not given such information (control condition). All participants were then presented with Alex's report. Depending on the condition, participants also either read that the report was prepared and processed by GrowthBar, a generative artificial intelligence (AI-disclosure condition) or were shown a spinning wheel (no-disclosure-control condition). After reading the report, participants were asked about the degree of trust they placed in Alex, based on the six-item cognition-based trust scale and the five-item affect-based trust scale (McAllister, 1995). An example item for affect-based trust is "I feel like I have a sharing relationship with Alex. We can both freely share our ideas, feelings, and hopes." When the items for cognition-based and affect-based trust were submitted to an exploratory factor analysis with iterated principal factors and orthogonal varimax rotation, we found two distinct factors, with cognition-based trust items clustering on one factor and affect-based trust items clustering on the other, indicating differentiation between the two constructs.

3.12.2. Results

Analysis of variance revealed a main effect of disclosure on trust, $F(1, 1044) = 65.67$, $p < 0.001$, $\eta^2 = 0.06$; no main effect of knowledge of AI usage on trust, $F(1, 1044) = 0.53$, $p = 0.47$, $\eta^2 = 0.001$; and an interaction effect of disclosure and knowledge of AI usage on trust, $F(1, 1044) = 3.48$, $p = 0.06$, $\eta^2 = 0.003$. In further support of H1, an independent-samples t test revealed that participants trusted Alex less when his report disclosed the usage of generative AI ($M = 4.10$, $SD = 1.15$) than when he made no such disclosure ($M = 4.62$, $SD = 0.92$; $t(1046) = 8.06$, $p < 0.001$, $d = 0.50$). Demonstrating the robustness of H1, the main effect of disclosure on trust was weaker—yet present—in the evaluator's-knowledge-of-AI-use-present condition ($d = 0.40$) compared to the control condition ($d = 0.60$). This finding demonstrates that the main effect of disclosure on trust is somewhat attenuated, but not eliminated, by prior knowledge of AI usage, indicating that the observed effect can be attributed primarily to the act of disclosure rather than to the mere fact of AI usage. While these analyses are based on the composite eleven-item trust measure, we also explored whether the main effect of AI disclosure on trust differs as a function of the type of trust being cognition-based versus affect-based. For this purpose, we ran

[5] Scholars such as Legood et al. (2023) have recently noted that McAllister's (1995) operationalization of trust overlaps considerably with the construct of trustworthiness from Mayer et al.'s (1995) model. Our intent here is not to weigh in on this debate but rather to explore whether our hypotheses are robust to differences in operationalization. Given that trustworthiness is arguably the most immediate predictor of trust (Colquitt et al., 2007), our results should be informative no matter whether McAllister's (1995) measure captures trust or trustworthiness.

seemingly unrelated regressions and conducted a Wald test to compare the coefficients. The results revealed a difference between the effects on cognition-based versus affect-based trust, $\chi^2(1) = 4.89$, $p = 0.03$, with AI disclosure having a comparatively stronger negative effect on cognition-based trust ($b =$ -0.58, $p < 0.001$) than on affect-based trust ($b =$ -0.44, $p < 0.001$).

3.13. Study 12

The aim of Study 12 was to test whether support for H1 remains robust regardless of whether the disclosure is voluntary or mandated by regulation. Disclosure regime (i.e., voluntary vs. mandatory) and the trustee's disclosure of AI usage were manipulated, and their effects on trust were measured. This 2 × 2 design allowed us to further disentangle the effect of AI disclosure from the effect of laws and regulations surrounding such disclosure.

3.13.1. Experimental procedure

We designed a task in which participants assumed the role of a consumer looking for a freelance graphic designer to design a postcard inviting guests to an upcoming dinner party. Before the main study, participants first read a news article about state legislatures moving towards regulating the disclosure of AI usage. Depending on the condition, the news article reported that state legislatures are moving towards either requiring disclosure (mandatory-disclosure-regime condition) or making disclosure voluntary (voluntary-disclosure-regime condition). Otherwise, the two versions of the article were almost identical in terms of content and length. Participants then learned that they had decided to hire the graphic designer Sebastian Belmont for the design of a postcard. Upon receiving an image of the postcard in their email inbox, and depending on the experimental condition, participants either read next to the postcard that it was designed with the help of Dall-E, generative artificial intelligence (AI-disclosure condition) or were not shown such information (no-disclosure-control condition). After seeing the postcard, participants were asked to rate the degree of trust they placed in Sebastian (adapted from Mayer & Davis, 1999) and the likelihood that they would hire Sebastian for another design job.

3.13.2. Results

Analysis of variance revealed a main effect of disclosure on trust, $F(1, 344) = 41.20$, $p < 0.001$, $\eta^2 = 0.11$; a main effect of disclosure regime on trust, $F(1, 344) = 10.41$, $p = 0.001$, $\eta^2 = 0.03$; and no interaction effect of disclosure and disclosure regime on trust, $F(1, 344) = 0.32$, $p = 0.57$, $\eta^2 = 0.001$. In further support of H1, an independent-samples t test revealed that participants trusted Sebastian less when he disclosed the usage of generative AI next to his postcard design ($M = 3.72$, $SD = 1.57$) than when he made no such disclosure ($M = 4.67$, $SD = 1.22$; across both disclosure-regime conditions, $t(346) = 6.31$, $p < 0.001$, $d = 0.68$). Subsample analyses showed that this main effect of disclosure on trust is present in both the mandatory-disclosure-regime condition ($d = 0.74$) and the voluntary-disclosure-regime condition ($d = 0.64$). Analysis of variance and t tests paint a highly similar picture for our secondary dependent variable, the likelihood to rehire the graphic designer. These findings suggest that the main effect of disclosure on trust persists regardless of whether regulations make disclosure voluntary or mandatory.

3.14. Study 13

The aim of Study 13 was to test H5 that an actor will be trusted least when exposed for (vs. disclosing and vs. not disclosing) having used AI for work tasks. Exposure or disclosure of AI usage by a tax advisor was manipulated and its impact on individual tax filers' trust in the tax advisor was measured.

3.14.1. Experimental procedure

During tax season, we asked participants to imagine they completed their income tax return using an online service company called Tax Return Service, which is highly similar to other such companies but with the benefit of inexpensively connecting filers with a human tax advisor. Participants were told they would engage in a chat session with a tax advisor named Charles Hewison. Participants were then connected in a chat window to Charles, who informed them how he planned to prepare their tax return by calculating their income and applying deductions and credits. After waiting for Charles to prepare the estimate of their expected tax refund or amount owed, participants were shown an estimated refund of $2,928 for the current tax year (which we had selected based on the average refund as determined by the IRS' filing statistics for the past five years from 2018 to 2022; IRS, 2024). Depending on the condition, participants either read that their tax refund estimate was prepared by generative AI (AI-disclosure condition); read that an article in their phone's newsfeed revealed that, according to an anonymous leak, Charles had prepared all of his tax refund estimates using generative AI (AI-exposure condition); or were shown a spinning wheel (no-disclosure-control condition). Subsequently, participants were asked to rate the degree of trust they placed in Charles (adapted from Mayer & Davis, 1999).

3.14.2. Results

Analysis of variance revealed an effect of disclosure on trust, $F(2, 192) = 37.41$, $p < 0.001$, $\eta^2 = 0.28$. In support of H5, pairwise-comparison t tests revealed that participants trusted Charles least when he was exposed for employing generative AI to estimate their tax refund ($M = 2.49$, $SD = 0.88$) compared to when he disclosed employing it ($M = 3.15$, $SD = 1.17$, $t(126) = 3.62$, $p < 0.001$, $d = 0.64$) or made no such disclosure ($M = 4.02$, $SD = 0.96$; $t(128) = 9.45$, $p < 0.001$, $d = 1.66$). In further support of H1, participants also trusted Charles less when he disclosed employing generative AI for estimating their tax refund than when he made no such disclosure ($t(130) = 4.65$, $p < 0.001$, $d = 0.81$).

3.15. Within-paper meta-analysis

The aims of our within-paper meta-analysis were (a) to synthesize the findings for the AI disclosure–trust main effect across our experiments, thereby providing a clearer, comprehensive estimate of the overall effect, and (b) to examine interactions between AI disclosure and four individual-level variables. Specifically, in each of our experiments, we collected information on each participant's level of (1) AI familiarity, (2) technology attitude, (3) AI usage, and (4) perceived AI accuracy. A meta-analytic approach is particularly appropriate for discovering relatively small effects that can be difficult to detect in individual studies (Goh et al., 2016), such as interaction effects. We posted the combined data and Stata syntax used in the within-paper meta-analysis to an OSF repository (see Table 1 for link).

3.15.1. Main effect of AI disclosure on trust

We used a random-effects approach as our default specification to account for heterogeneity across studies (Goh et al., 2016), but fixed-effects models produce substantively similar results, as shown in the log file included in the repository. Our meta-analysis focuses on the AI-disclosure-vs.-control contrast, so we omitted data from other conditions we collected in some studies as well as from Study 8 (in which we did not manipulate AI disclosure) from our meta-analytic data, resulting in a dataset comprising 4,093 observations. We also focused on perceptual trust (vs. behavioral or intention to trust) because we collected this information in all our experiments. Using the meta command in Stata, the meta-analysis confirmed that the level of trust was markedly lower in the AI-disclosure condition than in the control condition, $\theta = 0.81$, $z = 10.52$, 95 % CI [0.66, 0.96], $p < 0.001$. The *SM* includes a forest plot showing the estimates of the AI-disclosure-trust effect from the individual studies.

3.15.2. Interaction effects of AI disclosure and individual-level variables on trust

In each of our experiments, we measured AI familiarity using the question “How familiar are you with Artificial Intelligence and its applications (such as ChatGPT)?” (1 = *not at all familiar*; 5 = *very familiar*). We captured technology attitude with the question “What is your overall attitude towards new technological advancements?” (1 = *very negative*; 5 = *very positive*). Participants’ own AI usage was captured with the binary-choice question “Have you ever used AI tools or technologies in your professional work or business?” (*no*; *yes*). Finally, we gauged participants’ perception of AI accuracy by asking, “How reliable do you believe AI systems are in performing their tasks accurately?” (1 = *not reliable at all*; 5 = *highly reliable*).

To assess the moderating effect of each of these four variables on the influence of AI disclosure on trust, we ran ordinary least squares regressions with study number as clustering variable (Rabe-Hesketh & Skrondal, 2012) to account for differences across studies. Specifically, we ran four separate multivariate regression models with three predictors each, regressing trust on (1) AI disclosure, (2) one of the four individual-level variables, and (3) their interaction. In each of these four models, the main effect of AI disclosure on trust was negative and significant at $p < 0.001$. Further, we found positive main effects on trust for AI familiarity ($b = 0.04$, $z = 1.65$, 95 % CI [-0.01, 0.09], $p = 0.10$), technology attitude ($b = 0.19$, $z = 7.14$, 95 % CI [0.14, 0.24], $p < 0.001$), AI usage ($b = 0.13$, $z = 2.32$, 95 % CI [0.02, 0.24], $p = 0.02$), and perceived AI accuracy ($b = 0.26$, $z = 11.50$, 95 % CI [0.21, 0.30], $p < 0.001$), respectively. Turning to the individual moderating effects, the first model provided no compelling evidence that the AI disclosure × AI familiarity interaction predicts trust ($b = 0.00$, $z = 0.14$, 95 % CI [-0.06, 0.07], $p = 0.89$). In the model containing the AI disclosure × technology attitude interaction, this interaction term positively predicted trust ($b = 0.19$, $z = 5.38$, 95 % CI [0.12, 0.26], $p < 0.001$), indicating that the strength of the negative AI disclosure–trust effect diminishes among respondents with more positive attitudes towards technological advancements. Moreover, the AI disclosure × AI usage interaction had a positive effect on trust, but it failed to reach statistical significance by conventional standards ($b = 0.08$, $z = 1.04$, 95 % CI [−0.07, 0.22], $p = 0.30$).[6] Finally, we found a positive interaction effect of AI disclosure × perceived AI accuracy on trust ($b = 0.17$, $z = 5.35$, 95 % CI [0.11, 0.23], $p < 0.001$), such that the negative AI-disclosure effect of trust is attenuated if participants perceive that AI systems perform their tasks reliably. The *SM* shows the interaction plots for all four moderators. Further, region-of-significance plots revealed that the negative AI-disclosure effect was significant at $p < 0.05$ across the entire range of all moderators, suggesting that the moderators do not mute the focal main effect (also *SM*).

4. General Discussion

The purpose of this research was to examine what we term the AI-disclosure effect: that is, the influence of AI disclosure on trust across a diverse array of tasks. Across 13 experiments, we tested the theoretical prediction that an actor disclosing (vs. not disclosing) the usage of AI will be trusted less. The predicted effect was found not only in the general population but also among distinct professional cohorts, including legal analysts and hiring managers, as well as in student samples. Additionally, our work demonstrated that the focal effect manifests in various communication and writing tasks where generative AI is prevalently employed today, from mundane ones like composing an email to more significant ones like writing letters of application. Importantly, the observed effect extends beyond mere writing and communicative tasks to applications of AI in analytical functions (e.g., generating tax refund estimations) and artistic activities (e.g., advertising creation and graphic design). The robustness of this effect was also noted across different conceptualizations of independent and dependent variables, including both trust perceptions and behavioral intentions, such as the decision to hire a job applicant or to invest monetary funds. Additionally, the effect was found to be robust to the influence of several theoretically relevant variations, including the relative structural power of the trustor, the valence of the decision-making context, the way the disclosure is framed, the trustor’s knowledge of AI usage, and the disclosure regime. Finally, in our within-paper meta-analysis, we found that the negative effect of AI disclosure on trust is attenuated among people with positive attitudes towards technology and among those who perceive AI to be accurate; however, our analysis did not point to either AI familiarity or AI usage diminishing the negative effect.

4.1. Theoretical contributions

Our work makes several important contributions to both scholarly and public discourse. First, we contribute to the burgeoning organizational literature on the consequences of AI (Raisch & Fomina, forthcoming). Previous studies on the effects of AI usage at work have predominantly focused on operational impacts, such as enhanced productivity (Noy & Zhang, 2023), creativity (Jia et al., 2024), and decision-making quality (Gaessler & Piezunka, 2023). Our work shifts the narrative from focusing on operational outcomes toward appreciating the social outcomes of AI usage. We found that disclosing AI usage can negatively influence how trustworthy users are perceived to be, potentially affecting their career trajectories. This insight deepens our understanding of AI’s impact beyond mere task performance, emphasizing the social dimensions of technology usage in the workplace.

Second, we contribute to the literature on transparency (Bernstein, 2017). Our research challenges the prevailing assumptions that the effect of transparency is straightforwardly positive (Sah et al., 2018) and that transparency invariably builds trust (Schnackenberg & Tomlinson, 2016). Instead, our studies point to a paradoxical effect where disclosure of AI usage, intended to signal trustworthiness, ironically leads to reduced trust. By demonstrating the robustness of AI disclosure’s negative effect across various work settings and forms of AI usage, we shed new light on the delicate interplay between transparency and trust, suggesting that transparency may only be rewarded if the information conveyed is inherently unproblematic.

Third, we contribute to the trust literature, and more specifically, the research stream on trust-erosion processes (Guo et al., 2017). Prior findings indicate that trust erodes when expectations are not met (Elangovan et al., 2007; Lapidot et al., 2007). Building on micro-institutional theory (e.g., Bitektine & Haack, 2015; Zucker, 1977), our theorizing emphasizes the role of social norms in shaping such expectations and proposes legitimacy as a key mechanism explaining trust erosion. Our socio-cognitive model thus helps to answer calls for more research going beyond the trustor–trustee dyad to reveal how the social environment surrounding the dyad shapes trust (Gillespie et al., 2021). Additionally, our legitimacy account gives justice to the often neglected fact that trust judgments are regularly based on heuristic (rather than completely deliberate) considerations that stem from less conscious evaluations of a counterpart (Baer & Colquitt, 2018).

Fourth, we add to the growing body of scholarship at the intersection of trust and AI (Glikson & Woolley, 2020). Much of this work has started to investigate the extent to which humans trust AI (de Visser et al., 2016; Vanneste & Puranam, forthcoming). Our article joins recent conceptual

[6] To explore this effect further, we carried out a 2 × 2 experiment, in which we manipulated both the actual usage of a generative AI by study participants for writing a letter and, later in the study, the disclosure of AI usage by another person. We had expected that the AI-disclosure effect would be alleviated in the condition in which the study participants used AI themselves prior to evaluating the other person. However, what we instead found is that the AI-disclosure effect is remarkably robust. This study is reported as Study SM-3 in the *SM*.

work by ir250 to broaden this inquiry and investigate the extent to which humans trust other humans who use AI. That is, we study how AI affects trust in humans rather than in AI itself. This extension is important because it recognizes that trust in a technology can transfer to its human user, thus extending the literature on trust transfer (Stewart, 2003), which has traditionally focused on trust diffusion among humans.

4.2. Implications for managers, employees, and consumers

How can organizations decide between optional versus mandatory AI-disclosure policy, enforce it if needed, and foster an environment that maintains trust? Our findings inform organizations in strategizing about how they implement and communicate AI usage in the workplace, thus ensuring that such innovations hold the potential to enhance rather than undermine trust critical to the organization's success. More than one path exists for organizations to achieve this goal. They can make AI disclosure non-mandatory, thus protecting employees from the potential downsides that come with it, or they can consistently require all employees to disclose their AI usage. If they choose the latter route, we advise organizations to implement procedures that help them enforce AI disclosure—for instance, by routinely running work products through AI detectors. As our Study 13 shows, the threat of exposure could put the trust discount from disclosure into perspective, incentivizing employees to follow organizational disclosure policies. Concurrently, organizations may be well-advised to create an environment in which AI usage is perceived as collectively valid. In such settings, the trust-related consequences of disclosure are minimized (Study 7), thereby safeguarding individual employees from the negative effects of transparency.

How does the AI-disclosure effect inform the marketplace? The AI-disclosure effect revealed in our research has important implications not only for interactions within organizations but also for those in the marketplace, where trust functions as a critical foundation of consumers' responses to marketing (Morgan & Hunt, 1994). When consumers trust a brand, they are more likely to be loyal and engage in repeat purchasing, as trust reduces perceived risk and alleviates concerns about potential negative outcomes (Chaudhuri & Holbrook, 2001; Reimann et al., 2018). However, in light of our findings, disclosing the usage of AI in marketing efforts greatly diminishes consumer trust. Our studies—particularly, the one involving financial advertisements and the one on product design—provide direct support for this notion on at least two levels: trust in the content creator goes down and trust in the brand itself erodes. Because consumers are often closely attached to brands (Chaplin & Roedder John, 2005; Reimann, Castaño, Zaichkowsky, & Bechara, 2012; Reimann, Nuñez, & Castaño, 2017a), this reduced trust undermines key marketplace outcomes that it otherwise fosters: satisfaction, attachment, psychological ownership, and attention to and engagement with marketing efforts (Garbarino & Johnson, 1999; Thomson et al., 2005). Consequently, while AI may help marketers generate content faster and cheaper (Hartmann et al., 2024), transparency about AI involvement poses new challenges for maintaining the trust that is so vital to marketplace outcomes (Brüns & Meißner, 2024). Likewise, amidst potential benefits of AI for consumers (Puntoni et al., 2021), this research suggests that the very technology that enables these benefits also introduces distinct social and psychological costs, reduces perceived authenticity of product and service, makes the marketer and their brand seem less committed, and limits the overall value of AI-driven marketing.

How does the AI-disclosure effect compare to the privacy paradox? Both phenomena arise from concerns around technology, yet each addresses a fundamentally different issue. The privacy paradox refers to a gap between people's declared worries about data privacy and their surprisingly liberal data-sharing behaviors (Kokolakis, 2017). In contrast, the AI-disclosure effect highlights how individuals condemn others for using AI—even when doing so themselves, as revealed in our within-paper meta-analysis and Study SM-3. Thus, each phenomenon involves a contradiction between what people say and what they do, but the privacy paradox focuses on inconsistencies between one's own attitudes and actions, whereas the AI-disclosure effect involves interpersonal hypocrisy: judging others for behaviors one also engages in.

Isn't AI just a benign tool? Although people may rationalize their AI usage as "just a tool," our findings suggest that framing AI usage in this way does not protect against the trust erosion that disclosure can trigger (Study 9). Disclosing AI involvement in many shapes or forms can prompt audiences to wonder if technology is substituting for human effort, authenticity, and creativity. Therefore, while downplaying AI as a mere tool might seem tempting, our results indicate that trust-related concerns persist when people learn that AI has played a role, regardless of how it is labeled.

Would it be legitimate to disclose AI usage for ideation (rather than operations)? Researchers have argued that AI can act as a powerful partner in ideation (De Freitas et al., 2025), a stage of human creativity that precedes the operational tasks dominating our paper's empirical package (though our Studies 3 and 12 may be exceptions). However, we anticipate that the social and psychological costs stemming from negative reactions to AI-sourced ideation will be exceptionally high and need to be accounted for in appraising the overall value of AI usage for such purposes. We also wonder to what extent AI can truly ignite human creativity rather than constrain it through its focus on repackaging existing content, thus potentially commoditizing people's ideas rather than fostering genuinely novel breakthroughs.

4.3. Future directions

Our work has some limitations, which offer avenues for future research on the intricacies of AI disclosure. First, our investigation focuses on the evaluator's perspective but largely ignores the user's perspective, including the decision of whether or not to disclose AI usage. This dilemma often pits the virtue of transparency against the potential benefits of increased productivity and improved standing, creating a conflict between professional ethics and personal gain. More recent investigations point to the drive to disclose as an intrinsically rewarding force (Carbone & Loewenstein, 2023). Future research could explore how users reconcile such tensions when deciding to disclose AI usage. Understanding the processes that influence these decisions could help to identify factors that sway them in one direction or the other.

Second, the majority of our studies examine trust placed in previously unknown trustees—that is, swift trust (Blomqvist & Cook, 2018). Whether our findings extend to trust in established relationships remains an open question that should be addressed in future (possibly non-experimental) research.

Third, while our investigation covers a range of tasks, these primarily involve tasks that traditionally required considerable manual effort (see Study SM-1 in the *SM*). This limitation may affect the generalizability of our findings to highly automated settings. Future studies should broaden the scope to encompass automation-based tasks, thereby providing deeper insight into how AI disclosure affects trust across diverse work contexts.

Fourth, as workplaces increasingly adopt generative AI, the trust implications of AI disclosure identified in this investigation may evolve. Over time, with the expansion of AI in professional contexts and its normalization as a typical practice, reactions to AI disclosure could shift. While the interaction between AI disclosure and study participants' own usage of AI, which we began to explore in our meta-analysis and a supplemental study (see Study SM-3 in the *SM*), failed to reach statistical significance by conventional standards, the rapid evolution of generative AI, introducing capabilities that differ significantly from previous versions, may alter such dynamics. Future inquiries should therefore continue to investigate the impact of AI diffusion on trust erosion from AI disclosure, while also developing dynamic theories to explain the trajectories of AI legitimation over time. These inquiries will add to the body of scholarship on the inherently dynamic nature of trust (e.g., Aven et al., 2021).

Fifth, while we start to differentiate the effect of AI disclosure from that of algorithm aversion (Study 11), there is substantial room for additional investigation into this distinction. For example, future research could study how different contexts of AI disclosure (e.g., healthcare vs. financial services) differently impact perceptions of legitimacy and algorithm aversion, as well as how these perceptions sway trust. Moreover, future research could explore whether enhancements in legitimacy, made through communications about AI's ethical usage and alignment with professional standards, can mitigate the negative impact on trust more effectively than interventions designed to reduce algorithm aversion.

Sixth, our finding that third-party exposure of AI usage leads to even lower trust than self-disclosure (Study 13) presents interesting implications for better understanding trust dynamics related to AI usage. Future research can build on our results to venture deeper into why and how AI exposure and disclosure differ and the conditions under which this difference will be weaker or stronger. Such investigations will help identify the specific contexts in which the cost of disclosure will likely outweigh the damage resulting from potential third-party exposure.

Seventh, our study focuses on legitimacy as the mechanism underlying the AI-disclosure effect on trust, but this is not to say that no other mechanisms exist that could operate in parallel or in series with legitimacy. For instance, it could be insightful to consider Mayer et al.'s (1995) ability–benevolence–integrity model or Fiske et al.'s (2007) warmth–competence framework, which could usefully complement our focus on social appropriateness. Other research could go deeper and dissect legitimacy's mediating effect into its cognitive, moral, and pragmatic dimensions (Suchman, 1995).

5. Conclusion

This multi-study investigation illustrates that the seemingly straightforward and positive act of disclosing AI usage can substantially diminish trust across various professional settings from academia to business management. However, our findings also provide a roadmap for mitigating this trust deficit by showing that the negative impact of AI disclosure on trust is less pronounced among those with a favorable view of technology or positive perceptions of AI accuracy. Therefore, to foster trust in environments where AI is utilized, it is important to address workplace stakeholders' diverse attitudes towards technology. By tailoring communication strategies about AI applications to reflect these insights, organizations may better integrate AI into professional practices while maintaining trust among clients, colleagues, and collaborators. Overall, by offering a more detailed understanding of the factors that influence trust in the presence of AI, we equip organizations with the knowledge to better manage the social dynamics of technology use.

Author note

The authors thank the associate editor, Luke Zhu, and three anonymous reviewers, as well as Alex Bitektine, Bart de Jong, Patrick Haack, Nathan Podsakoff, and Bart Vanneste for helpful comments on earlier versions of this research. The authors are thankful for capable research assistance by Stephanie Clauson, Sydni Do, Christoph Hüller, and Amimul Zoha. Research support was provided by a National Endowment for the Arts research grant (#1925643–38-24) and a National Security Systems (TRIF NSS) research grant to the second author, and by a National Science Foundation CAREER Award (#1943688) to the first author. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This research was conducted in part while the authors were Visiting Faculty at EGADE Business School. The authors contributed equally to this article. Correspondence concerning this article should be addressed to Oliver Schilke, The University of Arizona, McClelland Hall 405GG, 1130 E. Helen St, Tucson, AZ 85721–0108. oschilke@arizona.edu.

Declaration of Generative AI and AI-assisted technologies in the writing process

Statement: During the preparation of this work, the authors used ChatGPT-4 and Dall-E (OpenAI, 2023) in order to generate the experimental stimuli (in particular the work products presented to participants, which are posted in OSF repositories linked within the manuscript), and to copyedit the manuscript. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Oliver Schilke: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization, Supervision. **Martin Reimann:** Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization, Software, Resources, and Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.obhdp.2025.104405.

Data availability

Data are linked in Table 1.

References

Agarwal, S. (2023). Overwhelmed by email? How AI tools can get you to inbox zero. *Wall Street J.* https://www.wsj.com/tech/ai/email-ai-tools-inbox-zero-72a8ac3f.

Aven, B., Morse, L., & Iorio, A. (2021). The valley of trust: The effect of relational strength on monitoring quality. *Organizational Behavior Human Decision Process., 166*, 179–193. https://doi.org/10.1016/j.obhdp.2019.07.004

Ali, A. E., Venkatraj, K. P., Morosoli, S., Naudts, L., Helberger, N., & Cesar, P. (2024). *Transparent AI disclosure obligations: Who, what, when, where, why, how Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613905.3650750

Baer, M. D., & Colquitt, J. A. (2018). Why do people trust? In R. H. Searle, A.-.-M.-I. Nienaber, & S. B. Sitkin (Eds.), *The Routledge companion to trust* (pp. 163–182). Routledge.

Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge*. Doubleday.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science, 58*(3), 739–753.

Bernstein, E. S. (2017). Making transparency transparent: The evolution of observation in management theory. *Academy of Management Annals, 11*(1), 217–266.

Birkinshaw, J., & Cable, D. (2017). The dark side of transparency. *McKinsey Quarterly, 1*, 88–95.

Bitektine, A., & Haack, P. (2015). The "macro" and the "micro" of legitimacy: Toward a multilevel theory of the legitimacy process. *Academy of Management Review, 40*(1), 49–75.

Blomqvist, K., & Cook, K. S. (2018). Swift trust: State-of-the-art and future research directions. In R. H. Searle, A.-.-M.-I. Nienaber, & S. B. Sitkin (Eds.), *The Routledge companion to trust* (pp. 29–49). Taylor & Francis.

Bovens, M. (2010). Two concepts of accountability: Accountability as a virtue and as a mechanism. *West European Politics, 33*(5), 946–967.

Brehm, J. W. (1966). *A theory of psychological reactance*. Academic Press.

Brüns, J. D., & Meißner, M. (2024). Do you create your content yourself? Using generative artificial intelligence for social media content creation diminishes perceived brand authenticity. *Journal of Retailing and Consumer Services, 79*, Article 103790.

Cañas, J. J. (2022). AI and ethics when human beings collaborate with AI agents. *Frontiers in Psychology, 13*, Article 836650.

Candelon, F., Krayer, L., Rajendran, S., & Zuluaga, D. M. (2023). How people can create—and destroy—value with generative AI. Retrieved March 16, 2024, from https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai.

Carbone, E., & Loewenstein, G. (2023). Privacy preferences and the drive to disclose. *Current Directions in Psychological Science, 32*(6), 508–514.

Chaplin, L. N., & Roedder John, D. (2005). The development of self-brand connections in children and adolescents. *Journal of Consumer Research, 32*(1), 119–129.

Chaudhuri, A., & Holbrook, M. B. (2001). The chain of effects from brand trust and brand affect to brand performance: The role of brand loyalty. *Journal of Marketing, 65*(2), 81–93.

Chen, S., Zhang, J. A., Gao, H., Yang, Z., & Mather, D. (2022). Trust erosion during industry-wide crises: The central role of consumer legitimacy judgement. *Journal of Business Ethics, 175*(1), 95–116.

Claeys, A.-S., Cauberghe, V., & Pandelaere, M. (2016). Is old news no news? The impact of self-disclosure by organizations in crisis. *Journal of Business Research, 69*(10), 3963–3970.

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology, 92*(4), 909–927.

Committee on Publication Ethics. (2023). Authorship and AI tools. Retrieved April 1, 2024, from https://publicationethics.org/cope-position-statements/ai-author.

De Freitas, J., Nave, G., & Puntoni, S. (2025). Ideation with generative AI. *Journal of Consumer Research.* in press.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied, 22*(3), 331–349.

Deephouse, D. L., Bundy, J., Tost, L. P., & Suchman, M. C. (2017). Organizational legitimacy: Six key questions. In R. Greenwood, C. Oliver, T. Lawrence, & R. E. Meyer (Eds.), *The SAGE handbook of organizational institutionalism* (2 ed., pp. 27–54). Sage.

Deephouse, D. L., & Carter, S. M. (2005). An examination of differences between organizational legitimacy and organizational reputation. *Journal of Management Studies, 42*(2), 329–360.

Deephouse, D. L., & Suchman, M. C. (2008). Legitimacy in organizational institutionalism. In R. Greenwood, C. Oliver, K. Sahlin, & R. Suddaby (Eds.), *The SAGE handbook of organizational institutionalism* (pp. 49–77). Sage.

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

Dirks, K. T., & de Jong, B. (2022). Trust within the workplace: A review of two waves of research and a glimpse of the third. *Annual Review of Organizational Psychology and Organizational Behavior, 9*(1), 247–276.

Elangovan, A. R., Auer-Rizzi, W., & Szabo, E. (2007). Why don't I trust you now? An attributional approach to erosion of trust. *Journal of Managerial Psychology, 22*(1), 4–24.

Elsbach, K. D. (1994). Managing organizational legitimacy in the California cattle industry: The construction and effectiveness of verbal accounts. *Administrative Science Quarterly, 39*(1), 57–88.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Management Unit Working Paper* (24-013).

Fishbowl. (2023). 70% of workers using chatgpt at work are not telling their boss; overall usage among professionals jumps to 43%. Retrieved April 14, 2024, from https://www.fishbowlapp.com/insights/70-percent-of-workers-using-chatgpt-at-work-are-not-telling-their-boss/.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83.

Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly, 77*(3), 515–540.

Gaessler, F., & Piezunka, H. (2023). Training with AI: Evidence from chess computers. *Strategic Management Journal, 44*(11), 2724–2750.

Garbarino, E., & Johnson, M. S. (1999). The different roles of satisfaction, trust, and commitment in customer relationships. *Journal of Marketing, 63*(2), 70–87.

Gay, R. (2024, May 17, 2024). If algorithms wrote your report, say so. *The New York Times*. https://www.nytimes.com/2024/03/16/business/work-friend-roxane-gay.html.

Gillespie, N., Fulmer, A., & Lewicki, R. (2021). A multilevel perspective on organizational trust. In N. Gillespie, A. Fulmer, & R. Lewicki (Eds.), *Understanding trust in organizations: a multilevel perspective* (pp. 3–13). Routledge.

Glaser, V. L., Fast, N. J., Harmon, D. J., & Green, S. (2016). Institutional frame switching: How institutional logics shape individual action. *Research in the Sociology of Organizations, 48A*, 35–69.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass, 10*(10), 535–549.

Gregory, A., & Ripski, M. B. (2008). Adolescent trust in teachers: Implications for behavior in the high school classroom. *School Psychology Review, 37*(3), 337–353.

Guo, S.-L., Lumineau, F., & Lewicki, R. J. (2017). Revisiting the foundations of organizational distrust. *Foundations and Trends in Management, 1*(1), 1–88.

Haack, P., Schilke, O., & Zucker, L. G. (2021). Legitimacy revisited: Disentangling propriety, validity, and consensus. *Journal of Management Studies, 58*(3), 749–781.

Harmon, D. J. (2019a). Arguments and institutions. *Research in the Sociology of Organizations, 65B*, 3–21.

Harmon, D. J. (2019b). When the Fed speaks: Arguments, emotions, and the microfoundations of institutions. *Administrative Science Quarterly, 64*(3), 542–575.

Hartmann, J., Exner, Y., & Domdey, S. (2024). The power of generative marketing: Can generative AI create superhuman visual marketing content? *International Journal of Research in Marketing, 42*(1), 13–31.

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford Press.

Holtz, B. C. (2015). From first impression to fairness perception: Investigating the impact of initial trustworthiness beliefs. *Personnel Psychology, 68*(3), 499–546.

IRS. (2024). *Filing season statistics - Individual income tax returns*. Retrieved March 21, 2024 from https://www.irs.gov/newsroom/filing-season-statistics-for-week-ending-april-21-2023.

Jia, N., Luo, X., Fang, Z., & Liao, C. (2024). When and how artificial intelligence augments employee creativity. *Academy of Management Journal, 67*(1), 5–32.

Johnson, C., Dowd, T. J., & Ridgeway, C. L. (2006). Legitimacy as a social process. *Annual Review of Sociology, 32*, 53–78.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors, 65*(2), 337–359.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals, 14*(1), 366–410.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology, 89*(1), 104–118.

Koehler, J. J., & Mercer, M. (2009). Selection neglect in mutual fund advertisements. *Management Science, 55*(7), 1107–1121.

Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & security, 64*, 122–134.

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology, 50*, 569–596.

Krausová, A., & Moravec, V. (2022). Disappearing authorship: Ethical protection of AI-generated news from the perspective of copyright and other laws. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law, 13*, 132.

Lapidot, Y., Kark, R., & Shamir, B. (2007). The impact of situational vulnerability on the development and erosion of followers' trust in their leader. *Leadership Quarterly, 18*(1), 16–34.

Lee, S. Y. (2016). Weathering the crisis: Effects of stealing thunder in crisis communication. *Public Relations Review, 42*(2), 336–344.

Legood, A., van der Werff, L., Lee, A., den Hartog, D., & van Knippenberg, D. (2023). A critical review of the conceptualization, operationalization, and empirical literature on cognition-based and affect-based trust. *Journal of Management Studies, 60*(2), 495–537.

Levine, S. S., Schilke, O., Kacperczyk, O., & Zucker, L. G. (2023). Primer for experimental methods in organization theory. *Organization Science, 34*(6), 1997–2025.

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of Management Review, 23*(3), 438–458.

Lockey, S., & Gillespie, N. (2024). Understanding trust in artificial intelligence: A research agenda. In R. C. Mayer, & B. M. Mayer (Eds.), *A research agenda for trust: Interdisciplinary perspectives* (pp. 11–24). Edward Elgar.

Loewenstein, G., Cain, D. M., & Sah, S. (2011). The limits of transparency: Pitfalls and potential of disclosing conflicts of interest. *American Economic Review, 101*(3), 423–428.

Loewenstein, G., Sah, S., & Cain, D. M. (2012). The unintended consequences of conflict of interest disclosure. *JAMA, 307*(7), 669–670.

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*(2), 90–103.

Lumineau, F., Schilke, O., & Wang, W. (2023). Organizational trust in the age of the Fourth Industrial Revolution: Shifts in the nature, production, and targets of trust. *Journal of Management Inquiry, 32*(1), 21–34.

Martin, K., & Waldman, A. (2023). Are algorithmic decisions legitimate? The effect of process and outcomes on perceptions of legitimacy of AI decisions. *Journal of Business Ethics, 183*(3), 653–670.

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology, 84*(1), 123–136.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal, 38*(1), 24–59.

McEvily, B., & Tortoriello, M. (2011). Measuring trust in organisational research: Review and recommendations. *Journal of Trust Research, 1*(1), 23–63.

Leavitt, K., Barnes, C. M., & Shapiro, D. L. (forthcoming). The role of human managers within algorithmic performance management systems: a process model of employee trust in managers through reflexivity. *Academy of Management Review*. https://doi.org/10.5465/amr.2022.0058.

McKinsey. (2023). *The state of AI in 2023: Generative AI's breakout year*. McKinsey & Company. Retrieved March 26, 2024 from https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-AIs-breakout-year.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems, 2*(2). Article 12.

Miller, D. T., & Morrison, K. R. (2009). Expressing deviant opinions: Believing you are in the majority helps. *Journal of Experimental Social Psychology, 45*(4), 740–747.

Morgan, R. M., & Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *Journal of Marketing, 58*(3), 20–38.

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*(5), 149–167.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192.

Ocasio, W. (2023). Institutions and their social construction: A cross-level perspective. *Organization Theory, 4*(3), Article 26317877231194368.

OECD. (2024). *Recommendation of the council on artificial intelligence*. OECD/LEGAL/0449.

OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat.

Palmeira, M., & Spassova, G. (2015). Consumer reactions to professionals who use decision aids. *European Journal of Marketing, 49*(3/4), 302–326.

Parsons, T. (1951). *The social system.* Free Press.

Podsakoff, P. M., & Podsakoff, N. P. (2019). Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *Leadership Quarterly, 30*(1), 11–33.

Polanyi, M. (1966). *The tacit dimension.* Doubleday.

Powell, W. W., & Colyvas, J. A. (2008). Microfoundations of institutional theory. In R. Greenwood, C. Oliver, K. Sahlin-Andersson, & R. Suddaby (Eds.), *The SAGE handbook of organizational institutionalism* (pp. 276–298). Sage.

Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing, 85*(1), 131–151.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata* (3rd ed.). Stata Press.

Raisch, S., & Fomina, K. (forthcoming). Combining human and artificial intelligence: hybrid problem-solving in organizations. *Academy of Management Review*. https://doi.org/10.5465/amr.2021.0421.

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

Reimann, M., Castaño, R., Zaichkowsky, J., & Bechara, A. (2012). How we relate to brands: Psychological and neurophysiological insights into consumer–brand relationships. *Journal of Consumer Psychology, 22*(1), 128–142.

Reimann, M., MacInnis, D. J., Folkes, V. S., Uhalde, A., & Pol, G. (2018). Insights into the experience of brand betrayal: From what people say and what the brain reveals. *Journal of the Association for Consumer Research, 3*(2), 240–254.

Reimann, M., Nuñez, S., & Castaño, R. (2017a). Brand-aid. *Journal of Consumer Research, 44*(3), 673–691.

Reimann, M., Schilke, O., & Cook, K. S. (2017b). Trust is heritable, whereas distrust is not. *Proceedings of the National Academy of Sciences, 114*(27), 7007–7012.

Sah, S. (2019). Conflict of interest disclosure as a reminder of professional norms: Clients first! *Organizational Behavior and Human Decision Processes, 154*(5), 62–79.

Sah, S., & Feiler, D. (2020). Conflict of interest disclosure with high-quality advice: The disclosure penalty and the altruistic signal. *Psychology, Public Policy, and Law, 26*(1), 88–104.

Sah, S., Malaviya, P., & Thompson, D. (2018). Conflict of interest disclosure as an expertise cue: Differential effects due to automatic versus deliberative processing. *Organizational Behavior and Human Decision Processes, 147*(4), 127–146.

Schilke, O. (2018). A micro-institutional inquiry into resistance to environmental pressures. *Academy of Management Journal, 61*(4), 1431–1466.

Schilke, O., & Lumineau, F. (forthcoming). How organizational is interorganizational trust? How organizational is interorganizational trust? Academy of Management Review. https://doi.org/10.5465/amr.2022.0040.

Schilke, O., Powell, A., & Schweitzer, M. E. (2023). A review of experimental research on organizational trust. *Journal of Trust Research, 13*(2), 102–139.

Schilke, O., Reimann, M., & Cook, K. S. (2015). Power decreases trust in social exchange. *Proceedings of the National Academy of Sciences, 112*(42), 12950–12955.

Schilke, O., Reimann, M., & Cook, K. S. (2021). Trust in social relations. *Annual Review of Sociology, 47*, 239–259.

Schilke, O., & Rossman, G. (2018). It's only wrong if it's transactional: Moral perceptions of obfuscated exchange. *American Sociological Review, 83*(6), 1079–1107.

Schilke, O., Xue, Z., & Haack, P. (forthcoming). Legitimacy construction in the presence of multiple validity cues: an experimental investigation. In J. E. Stets, K. A. Hegtvedt, & L. Doan (Eds.), *Handbook of social psychology: micro, meso, and macro orientations*. Oxford University Press.

Schnackenberg, A. K., & Tomlinson, E. C. (2016). Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships. *Journal of Management, 42*(7), 1784–1810.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89*(6), 845–851.

Stewart, K. J. (2003). Trust transfer on the world wide web. *Organization Science, 14*(1), 5–17.

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review, 20*(3), 571–610.

Suddaby, R., Bitektine, A., & Haack, P. (2017). Legitimacy. *Academy of Management Annals, 11*(1), 451–478.

Thomson, M., MacInnis, D. J., & Whan Park, C. (2005). The ties that bind: Measuring the strength of consumers' emotional attachments to brands. *Journal of Consumer Psychology, 15*(1), 77–91.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge University Press.

Treviño, L. K., den Nieuwenboer, N. A., Kreiner, G. E., & Bishop, D. G. (2014). Legitimating the legitimate: A grounded theory study of legitimacy work among Ethics and Compliance Officers. *Organizational Behavior and Human Decision Processes, 123*(2), 186–205.

Vanneste, B. S., & Puranam, P. (forthcoming). Artificial intelligence, trust, and perceptions of agency. *Academy of Management Review*.

Zelditch, M., & Walker, H. A. (1984). Legitimacy and the stability of authority. *Advances in Group Processes, 1*, 1–25.

Zetwerk. (2024). Should businesses disclose their AI usage? Retrieved April 14, 2024, from https://www.zetwerk.com/ai-disclosure-in-business/.

Zucker, L. G. (1977). The role of institutionalization in cultural persistence. *American Sociological Review, 42*(5), 726–743.

Zucker, L. G. (1987). Institutional theories of organization. *Annual Review of Sociology, 13*, 443–464.

Zucker, L. G., & Schilke, O. (2019). Towards a theory of micro-institutional processes: Forgotten roots, links to social-psychological research, and new ideas. *Research in the Sociology of Organizations, 65B*, 371–389.