Journal of Retailing and Consumer Services 94 (2027) 105001

Contents lists available at ScienceDirect

Journal of Retailing and Consumer Services

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

No one is watching: AI service agents and the rise of unethical consumer behavior

Lélis Balestrin Espartel [a,b], Simoni F. Rohden [c,d,*], Clécio Falcão Araújo [e]

[a] IADE – Universidade Europeia, Estrada da Correia, 53, Lisbon, 1500-210, Portugal
[b] Center for Transdisciplinary Development Studies (CETRAD-Europeia), Portugal
[c] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, Lisbon, 1070-312, Portugal
[d] Information Management Research Center (MagIC), Portugal
[e] PUCRS Business School, Pontifícia Universidade Católica do Rio Grande do Sul, Av. Ipiranga, 6681, Porto Alegre, RS, Brazil

ARTICLE INFO

Keywords:
Social judgment
Moral self-regulation
Consumer ethics
AI-Mediated interactions
Moral identity
Responsible retailing

ABSTRACT

The growing use of artificial intelligence (AI) agents in frontline service roles is reshaping frontline interactions, with important implications for consumer ethical behavior. This research examines whether AI-mediated service interactions increase unethical consumer behavior and identifies the psychological mechanism underlying this effect. Across a pilot study and three experiments, we show that consumers who interact with AI agents report higher unethical behavioral intentions than those who interact with human employees. This effect is explained by reduced perceived social judgment, defined as the extent to which consumers perceive the interacting agent as capable of morally evaluating their behavior, rather than by entitlement, anticipatory guilt, or moral disengagement. The effect is strongest when both internal (moral identity) and external (detectability) accountability are low. Importantly, cognitive priming of evaluative concern does not restore this perception in AI-mediated interactions, suggesting that reduced social judgment is anchored in the agent's perceived evaluative capacity rather than in momentary cognitive salience. These findings show that AI alters the evaluative dynamics of retail and service encounters, increasing the risk of unethical consumer behavior, and highlight the need for firms to embed accountability mechanisms when deploying AI in customer-facing roles.

1. Introduction

Artificial intelligence (AI) is rapidly reshaping consumer purchase experiences (Klaus and Zaichkowsky, 2022; Rohden and Espartel, 2026). Across retail and service settings, firms increasingly rely on AI-powered solutions to improve efficiency, scalability, and service accessibility (Brüggemann et al., 2025; Schultz and Paetz, 2025; Tran, 2026). However, replacing human employees with AI agents raises concerns that extend beyond operational performance. Service encounters frequently depend on consumer honesty, for example, when reporting product issues or resolving billing discrepancies (Reynolds and Harris, 2009; Tomazelli et al., 2024). When honesty breaks down, consumers may engage in unethical or deviant behaviors, including misreporting information or acting abusively toward service providers (Cheng, 2025; Lei et al., 2025; Neale and Fullerton, 2010). In such contexts, the nature of the interacting agent becomes consequential, shaping not only the service experience but also consumers' ethical decision-making (Sohn et al., 2025).

In service encounters, human social presence activates reputational concerns and self-monitoring (Tangney et al., 2007). By contrast, AI agents are often perceived as lacking consciousness, moral agency, and the capacity to form genuine social judgments (Myers and Everett, 2025; Zhao et al., 2025). Consumers may therefore feel less judged and less morally constrained when interacting with AI rather than human employees, increasing the likelihood of opportunistic or irresponsible consumer decisions during marketplace interactions. This raises an important challenge for responsible retailing and service management, as AI-enabled service environments may inadvertently weaken the social and moral safeguards that support ethical consumer behavior (Holtrop et al., 2025). As firms increasingly replace or supplement

This article is part of a special issue entitled: Digital & Responsible Retailing published in Journal of Retailing and Consumer Services.
The article's data will be shared upon request with the corresponding author.

* Corresponding author. NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312, Lisbon, Portugal.
E-mail addresses: lelis.espartel@universidadeeuropeia.pt (L.B. Espartel), srohden@novaims.unl.pt (S.F. Rohden), clecio.araujo@pucrs.br (C.F. Araújo).

https://doi.org/10.1016/j.jretconser.2026.105001
Received 11 May 2026; Received in revised form 22 July 2026; Accepted 22 July 2026
Available online 25 July 2026
0969-6989/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

frontline employees with AI agents, understanding how service design shapes ethical consumer decision-making becomes a critical issue for retailing and consumer services research.

Despite growing evidence that AI agents influence unethical consumer behavior, prior research has focused on internal psychological mechanisms, such as reduced anticipatory guilt (Kim et al., 2023), moral disengagement (Cheng, 2025), and entitlement (Cao et al., 2023), while paying limited attention to how the design of the retail or service interaction itself, particularly the absence of a human evaluator, shapes these effects. Moreover, studies have examined individual and contextual factors in isolation (Lei et al., 2025; Dootson et al., 2023), without testing how internal (e.g., moral identity) and external (e.g., detectability) constraints jointly influence unethical behavior.

Building on these gaps, this research proposes that AI agents increase unethical consumer behavior by reducing perceived social judgment, defined as the extent to which consumers feel evaluated by an entity capable of moral appraisal (Kim et al., 2022). We argue that, unlike frontline employees, AI agents do not constitute credible moral audiences, weakening evaluative cues that typically reinforce moral restraint. Importantly, this reduction reflects the agent's perceived evaluative capacity, a structural feature of the interaction, rather than a temporary cognitive state. As a result, it shapes whether internal regulatory processes are activated, systematically influencing the decision environment in which consumers evaluate and act on ethically relevant choices.

This research makes three theoretical contributions. First, it identifies perceived social judgment as a key mechanism through which AI-mediated retail and service interactions influence unethical consumer behavior, shifting attention from internal moral states to the social-evaluative structure of customer-facing encounters. Second, it shows that this mechanism is anchored in the agent's perceived evaluative capacity and is not restored by cognitive salience alone, while remaining contingent on agent design features (e.g., anthropomorphism and hybrid human-AI configurations), highlighting the limits of managerial interventions aimed at promoting responsible consumer behavior in AI-enabled services. Third, it demonstrates that this effect is jointly constrained by moral identity and detectability, clarifying how internal and external accountability mechanisms can partially compensate for the absence of human evaluation, offering insights for the design of more responsible retail and service environments.

2. Literature review and hypothesis development

The increasing deployment of AI agents in retail and service environments is transforming the moral dynamics of consumer-firm interactions. Service encounters are inherently social exchanges; substituting human employees with AI may alter the social architecture of the interaction itself, weakening the structural cues that typically activate moral constraint in marketplace behavior. Unlike human frontline employees, AI agents are not perceived as intentional moral evaluators capable of forming social judgments or holding consumers accountable (Myers and Everett, 2025; Zhao et al., 2025). While prior research has emphasized the efficiency and personalization benefits of AI in services (Tran, 2026), less attention has been paid to how replacing human agents may alter the psychological mechanisms that regulate consumer misconduct. Existing literature has mostly focused on mechanisms such as anticipatory guilt, moral disengagement, and perceptions of AI humanness (Table 1). Closer to our focus, Li et al. (2024) also identify perceived social judgment as a mediator, linking it to AI identity disclosure in a single retail-fraud scenario. We build on this construct but extend its theoretical role in three ways: by testing it as the mechanism linking agent type (AI vs. human) rather than disclosure to unethical behavior; by specifying the joint boundary conditions under which the indirect effect operates (moral identity and detectability as substitutable accountability mechanisms); and by experimentally testing whether the construct is cognitively malleable or structurally

Table 1
Literature on the influence of AI interaction on unethical consumer behavior.

Paper	Context	IV	DV	Mediators	Moderators	Other variables
Giroux et al. (2022)	Service (restaurant) Retail store	Service agent (human vs. Machine vs. AI)	Moral intention	Anticipatory guilt	AI perception (humanlike vs. Non humanlike)	Empathy, preventability, detectability, type of mistake, and honest mistake
Kim et al. (2023)	Lottery Services (tutoring) Retail (fashion)	Service agent (human vs. robot)	Unethical behavior	Anticipatory guilt	Type of agent (functional vs. Social)	Service engagement, choice confidence, warmth, detectability, empathy
Liu et al. (2023)	Service (restaurant, hotel)	Service agent (human vs. robot) Service exclusion (inclusion vs. exclusion)	Unethical behavior	Anticipatory guilt	Anthropomorphism	Warmth, competence, detectability, empathy, belonging, control, self-esteem, sense of meaning
Dootson et al. (2023)	Service (ATM)	Humanness of service agent (low vs. High)	Opportunistic deviance	Perceived risk Perceived empathy	Attitude toward robots	-
Li et al. (2024)	Retail (fashion) Services (restaurant, insurance)	AI identity (disclosure vs. non-disclosure)	Unethical behavior	Social judgment	Impression management motive	Warmth, empathy, skepticism, self-awareness
Lei et al. (2025)	Retail (fashion, mobile phones) Services (insurance)	Role of AI agent (servant vs. partner)	Unethical behavior	Moral disengagement	Moral identity AI appearance Visibility	-
Zhu et al. (2025)	Services (insurance, restaurant)	Service agent (human vs. AI)	Unethical behavior	Anticipatory guilt Responsibility attribution	Social distance	Anticipatory shame
Deng et al. (2026)	Retail (supermarket) Services (restaurant)	Degree of robot adoption (human vs. co-delivered vs. Robot)	Unethical behavior	Social pressure Perceived electronic monitoring Perceived entitativity	Social rewards-punishments	Anticipatory guilt, mind perception
This research	Services (airline) Retail (computer)	Agent type (human vs. AI)	Unethical behavior	Social judgment	Moral identity Detectability	Entitlement, self-control, moral disengagement, AI familiarity, AI avoidance

anchored in the agent's perceived evaluative capacity. This positions our contribution as an extension of the construct's theoretical role and boundary conditions.

From a moral self-regulation perspective, ethical behavior is maintained through ongoing self-monitoring processes in which individuals evaluate their actions against internalized moral standards (Bandura, 1991; Aquino and Reed, 2002). These self-regulatory processes are often triggered by social-evaluative cues, which signal that one's behavior is being observed or judged (Waqas and Qadri, 2025). The presence of another person can heighten reputational concerns and trigger self-conscious emotions, such as guilt or shame, which inhibit unethical conduct (Tangney et al., 2007).

Interactions with AI agents may attenuate these moral self-regulatory triggers because AI lacks perceived consciousness, moral agency, and the capacity for social judgment. Consumers may feel less evaluated and less morally accountable during AI-mediated exchanges (Li et al., 2024; Lei et al., 2025). Emerging evidence suggests that AI interactions can reduce anticipatory guilt and increase willingness to engage in unethical consumer behaviors (Kim et al., 2023). In retail contexts, where opportunities for unethical behaviors such as misreporting, exaggerating claims, or policy exploitation are common, diminished activation of moral self-regulation may make such behaviors more psychologically permissible (Tomazelli et al., 2024). Accordingly, we expect that when interacting with AI rather than human agents, consumers may experience weaker activation of moral self-regulatory processes, leading to greater intentions to engage in unethical behaviors. Therefore, we hypothesize.

H1. Interacting with an AI agent (vs. a human employee) increases consumers' intention to engage in unethical behavior.

Although AI agents may increase unethical behavior, understanding why this occurs requires identifying the psychological mechanism through which AI interactions alter moral self-regulation. Moral self-regulation theories suggest that ethical conduct is sustained through ongoing self-monitoring against internalized moral standards (Bandura, 1991; Waqas and Qadri, 2025). Critically, this self-monitoring is often triggered by evaluative cues that signal the presence of a plausible human audience capable of observing and morally evaluating one's behavior. When individuals perceive that another party can evaluate their actions, moral standards become more salient, thereby strengthening behavioral restraint (Li et al., 2024).

Human employees naturally provide such evaluative cues. Their perceived intentionality, emotional capacity, and moral agency make them credible evaluators of consumer behavior. Research shows that the presence of another human can heighten norm adherence and reduce unethical behavior due to reputational and self-presentational concerns (Argo et al., 2005). In contrast, AI agents are typically perceived as lacking consciousness and moral awareness, thereby reducing the extent to which consumers feel judged or morally evaluated during interactions (Kim et al., 2022; Zhao et al., 2025). Consequently, AI-mediated encounters may attenuate perceived social judgment not merely by reducing momentary salience of evaluation, but also by reducing the agent's perceived evaluative capacity, thereby weakening downstream moral restraint (Li et al., 2024).

When perceived social judgment is reduced, moral self-regulatory processes may weaken. Without the psychological trigger of evaluation, consumers may experience diminished anticipatory guilt and reduced concern about maintaining a moral self-image (Kim et al., 2023; Tangney et al., 2007). We argue that interacting with an AI agent decreases perceived social judgment, which in turn reduces the activation of moral self-regulation and increases unethical behavioral intentions. Therefore.

H2. The effect of interacting with an AI agent (vs. a human employee) on unethical behavior is mediated by reduced perceptions of social judgment.

Consumers differ in how strongly social-evaluative cues regulate their behavior. Moral identity reflects the extent to which a person's self-concept is rooted in moral domains (Aquino and Reed, 2002; Sreen et al., 2025). Individuals high in moral identity are more motivated to maintain a moral self-concept and to avoid actions that would threaten who they are, effectively serving as an internal accountability constraint when external evaluative cues are weak (Zhao et al., 2025). Consequently, their restraint is anchored internally and does not depend on feeling evaluated by an external audience, allowing moral identity to compensate for reduced external accountability in AI-mediated interactions (Lei et al., 2025). In other words, when moral identity is strong, behavior is governed by internalized standards regardless of perceived social judgment, so variation in social-evaluative cues has little downstream effect; when moral identity is weak, behavior is contingent on external evaluation, so reduced perceived social judgment translates more directly into unethical intentions. This aligns with broader consumer-morality perspectives, emphasizing that moral motives and identity-based concerns shape when moral standards guide marketplace behavior. Previous literature has suggested that high moral identity reduces consumers' (Lei et al., 2025) and employees' (Zhao et al., 2025) propensity to engage in unethical behavior when interacting with AI agents.

Furthermore, transgression detectability may play a relevant role in this context. Fraud and unethical behavior are frequent not only because of their financial returns, but also because they are often undetectable (Tergiman and Villeval, 2022). Detectability captures an external accountability constraint that reinstates the possibility of reputational evaluation and sanction, thereby compensating for reduced social judgment in AI-mediated interactions (Dootson et al., 2023). When misconduct is likely to be detected, concerns about reputation and potential sanctions become salient, restoring the behavioral relevance of social evaluation (Deng et al., 2026).

Although moral identity and detectability are conceptually independent (one dispositional, the other a situational property of the transgression), a robust regularity across the morality–deterrence literature indicates that they operate as functional substitutes in regulating misconduct. Internalized moral standards can render external sanctions and detection cues largely irrelevant, such that, when moral commitments are strong, instrumental considerations of being caught carry little weight, and they become consequential primarily when internal restraint is weak (Paternoster and Simpson, 1996; Kroneberg et al., 2010). The same internal–external distinction structures behavioral ethics accounts of consumer dishonesty (Mazar et al., 2008; Ayal et al., 2015). Consequently, the behavioral relevance of detectability should be strongest when moral identity is low and attenuated when moral identity is high. This substitution logic implies a joint moderation rather than two independent moderating effects. Extending this logic to AI-mediated service encounters, the reduced perceived evaluative capacity removes the default social-evaluative constraint, leaving moral identity (internal) and detectability (external) as the remaining, mutually substitutable safeguards. Because either is sufficient to reinstate restraint, the amplifying effect of reduced social judgment on unethical behavior should emerge specifically when both are absent, indicating a multiplicative rather than additive joint moderation.

In this sense, we expect the indirect effect of agent type (AI vs. human) on unethical behavior via reduced social judgment to be strongest under conditions of low moral identity (weak internal regulation) and low transgression detectability (weak external accountability). Conversely, when either moral identity is high (strong internal accountability) or detectability is high (strong external accountability), the downstream impact of reduced social judgment should be attenuated. In these situations, at least one accountability structure remains operative, serving as a substitute for the missing human evaluative presence and limiting unethical intentions (Kim et al., 2023). Therefore, we propose that.

H3. Internal (moral identity) and external (detectability) accountability mechanisms function as substitutes in constraining unethical behavior. Accordingly, the indirect effect of agent type on unethical behavioral intentions through perceived social judgment will be strongest when both moral identity and detectability are low and weaker when either moral identity or detectability is high.

Fig. 1 summarizes the proposed conceptual model linking agent type, perceived social judgment, moral identity, detectability, and unethical consumer behavior.

2.1. Overview of studies

We conducted a series of studies to examine whether and why interactions with AI agents increase unethical consumer behavior relative to human interactions. A pilot study provided initial evidence of greater tolerance for unethical conduct. Study 1 established a main effect and showed that reduced perceived social judgment mediates this relationship, ruling out entitlement as a mediator. Study 2 identified boundary conditions, demonstrating that the effect depends on detectability and moral identity, while ruling out the role of moral disengagement. Study 3 tested whether social judgment is cognitively malleable or anchored in the agent's evaluative capacity, providing evidence consistent with the latter explanation.

3. Pilot study

Service interactions are inherently social exchanges. When consumers interact with a human employee, the presence of another person may activate reputational concerns and perceived social evaluation, thereby constraining unethical behavior (Lei et al., 2025). In contrast, AI agents are often perceived as socially neutral and non-evaluative, which may weaken such moral restraint (Li et al., 2024). As an initial test of this proposition, we examine whether interacting with an AI agent (vs. a human employee) increases tolerance for unethical consumer behavior in a service recovery context (Cheng, 2025). We predict that participants who interact with an AI agent will report greater acceptance of unethical consumer conduct than those who interact with a human employee.

3.1. Procedures

We conducted a randomized between-subjects experiment to test the effect of agent type (human vs. AI) on tolerance for unethical behavior in a service recovery context. After applying the study exclusion criteria, the final sample consisted of 94 valid participants. Participants read a scenario involving an airline reimbursement for lost luggage. Sam's luggage was lost, and although the contents were valued at £500, the airline's online reimbursement system allowed claims of up to £1500 without requiring proof of the contents. The ethical dilemma concerned whether Sam should report the actual loss or overstate the reimbursement amount. The detailed scenarios are available in Appendix 1. A projective approach (i.e., evaluating Sam's behavior) was used to reduce social desirability bias and elicit more candid judgments about unethical conduct. The manipulation varied depending on whether a human employee or an AI agent processed the claim. Participants were recruited via Prolific and were UK residents fluent in English. The same recruitment platform and eligibility criteria were used across studies.

3.2. Measurements

After providing informed consent and reading the scenario, participants answered one question about intentions to engage in unethical behavior: "How much money do you think Sam will ask the company?" (adapted from Neale and Fullerton, 2010). The answer was provided with a slider ranging from £0 to £1500. Additional measurements included participant familiarity with AI (adapted from Rohden and Espartel, 2026) and AI avoidance ("Whenever possible, I avoid using Artificial Intelligence technologies") to control for potential confounds related to individuals who use AI tools less often. Demographic questions included age and gender.

3.3. Results

Participants were evenly distributed across conditions (50% female; M_{age} = 43 years, SD = 14.61), with 49 participants assigned to the human condition and 45 to the AI condition. There were no significant differences across conditions in familiarity with AI ($F(1, 92) = .06$, $p = .814$) or AI avoidance ($F(1, 92) = .04$, $p = .852$); therefore, these variables were not included as covariates.

An Analysis of Variance (ANOVA) was conducted to examine the effect of agent type on the compensation amount participants believed Sam should request, and the results revealed a significant effect of agent type ($F(1, 92) = 4.24$, $p = .042$, $\eta^2 = .044$). Participants exposed to the human employee condition suggested an average reimbursement of £726.14 (SD = 185.26), whereas participants in the AI condition endorsed a significantly higher amount of £807.49 (SD = 197.57). Thus, although participants in both conditions believed some overstatement was acceptable, those in the AI condition supported greater inflation of the reimbursement claim.

3.4. Discussion

The pilot study provides initial evidence that interactions with AI agents may loosen moral boundaries in consumer decision-making (Cao et al., 2023; Kim et al., 2023). Participants reported higher compensation claims when interacting with an AI agent than when interacting

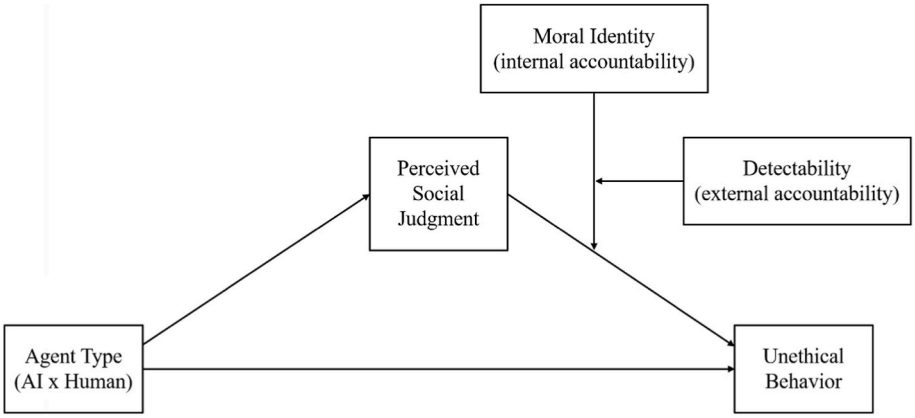

Fig. 1. Conceptual model of agent service interactions and unethical consumer behavior.

with a human employee. However, the study has limitations: it relied on a projective method, which may reduce personal moral engagement (Dootson et al., 2023), and the scenario involved a prior service failure, potentially confounding results with retaliatory motives. These limitations motivated the subsequent studies, which used self-referential decisions to test the proposed mechanism while eliminating the service-failure confound.

4. Study 1

Study 1 builds on pilot evidence by examining the psychological mechanism underlying the effect. We propose that the key difference between AI and human service interactions lies in perceived social evaluation (Li et al., 2024). Human presence activates evaluative concerns that reinforce moral self-regulation (Lei et al., 2025), whereas AI agents are not seen as capable of moral judgment (Dong et al., 2025), reducing perceived social judgment. Accordingly, we predict that AI interactions increase unethical intentions, with this effect mediated by lower perceived social judgment. Study 1, therefore, isolates whether the mere absence of a human evaluator, rather than the disclosure of an agent's identity (Li et al., 2024), is sufficient to reduce perceived social judgment and increase unethical intentions.

4.1. Procedures

Study 1 employed a randomized between-subjects, single-factor design to examine the impact of agent type (human vs. AI) on unethical behavioral intentions. After applying the study exclusion criteria, the final sample consisted of 128 valid participants. Participants were asked to imagine purchasing a computer online for £1000 and applying a 10% discount voucher. Due to a system error, they received a 30% discount instead. The ethical dilemma involved deciding whether to report the mistake or retain the additional £200. In the human condition, participants interacted with "Alex," a company employee. In the AI condition, the interaction occurred with an AI agent available on the company's website.

4.2. Measurements

After providing informed consent and reading the scenario, participants answered four items measuring unethical behavior intentions (e.g., "In this situation I would keep the extra £200", adapted from Schaefers et al., 2016). Perceived social judgment was measured using three items (e.g., "I would be anxious that the service agent might have a bad impression of me", adapted from Kim et al., 2022). As control variables, we measured self-control with four items and psychological entitlement with nine items, both adapted from Campbell et al. (2004). Self-control was included because it is associated with ethical decision-making and behavioral restraint, whereas psychological entitlement was measured to rule out an alternative explanatory mechanism previously linked to unethical behavior (Cao et al., 2023). Familiarity with AI and AI avoidance (1 item each) were included because prior experience with and attitudes toward AI may influence consumers' responses to AI-mediated interactions independently of the experimental manipulation. The full list of items and scales is available in Appendix 2.

As a manipulation check, participants indicated their agreement with the statement "Customer service was operated by an artificial intelligence". All measures were assessed with 7-point Likert scales (ranging from "strongly disagree" to "strongly agree"). In addition to the multi-item scales, participants also completed one nominal attention check ("What product was described in the scenario?") and provided demographic information (gender and age).

4.3. Results

The resulting sample comprised 128 respondents (51% female; M_{age} = 46 years, SD = 12.96). The manipulation check confirmed that participants correctly perceived the agent type ($F(1, 126) = 701.65$, $p < .001$). Participants in the human condition strongly disagreed that the interaction involved AI ($M_{disagreed} = 1.55$, $SD = 1.48$), whereas those in the AI condition strongly agreed ($M_{agreed} = 6.89$, $SD = .65$).

The inclusion of familiarity with AI, AI avoidance, and self-control did not alter the substantive pattern of results. Therefore, for parsimony, we report the analyses without these covariates. An ANOVA revealed a significant main effect of agent type on unethical intentions ($F(1, 126) = 3.99$, $p = .048$). Participants who interacted with the AI agent reported a greater propensity to keep the extra £200 ($M_{AI} = 5.41$, $SD = 1.67$) than those who interacted with a human employee ($M_{Human} = 4.74$, $SD = 2.06$).

To test the proposed explanatory mechanism, we conducted a mediation analysis using PROCESS Model 4 (Hayes, 2017). As predicted, agent type significantly influenced perceptions of social judgment (path a), such that interacting with an AI agent reduced perceived social judgment compared to interacting with a human employee ($\beta = -.65$, $t = -2.92$, $p = .004$). In turn, lower perceptions of social judgment were associated with stronger intentions to behave unethically (path b; $\beta = -.54$, $t = -4.38$, $p < .001$).

Importantly, once the mediator was included in the model, the direct effect of agent type on unethical intentions became non-significant ($\beta = .31$, $t = .97$, $p = .34$), while the indirect effect remained significant ($\beta = .35$, $BootSE = .14$, 95% CI [.109, .637]). This pattern indicates that reduced perceived social judgment fully accounts for the effect of AI interaction on unethical behavioral intentions. In other words, consumers interacting with an AI agent feel less socially evaluated, which in turn increases their likelihood of engaging in unethical conduct.

We further examined whether entitlement could serve as an alternative explanatory mechanism. Although feelings of entitlement were positively associated with unethical intentions ($\beta = .31$, $t = 2.52$, $p = .013$), agent type did not significantly affect entitlement ($\beta = -.30$, $t = -1.29$, $p = .199$). Consistent with this, the indirect effect through entitlement was not significant ($\beta = -.05$, $BootSE = .04$, 95% CI [-.145, .026]). These findings rule out entitlement as an alternative mediating pathway, strengthening the conclusion that reduced social judgment is the primary mechanism driving the effect.

4.4. Discussion

Study 1 provides causal evidence that AI-agent interactions increase unethical intentions, as participants who interacted with AI reported a greater willingness to behave unethically than those who interacted with human employees, hence supporting H1. Moreover, social aspects influence how consumers experience and respond to interactions with technology in the purchase journey (Schultz and Paetz, 2025). The enhanced unethical behavior is driven by reduced perceptions of social judgment (Li et al., 2024), which fully mediate the relationship, ruling out alternative explanations such as entitlement (Li et al., 2024; Sohn et al., 2025), confirming H2.

However, the study does not account for contextual and individual factors that may constrain this effect. Specifically, it remains unclear whether increased detectability can restore accountability (Dootson et al., 2023) or whether differences in moral identity influence sensitivity to reduced social judgment (Zhao et al., 2025). These limitations motivated Study 2, which examines the joint role of detectability and moral identity as boundary conditions (Lei et al., 2025; Li et al., 2024).

5. Study 2

Theoretically, Study 2 tests whether this reduced evaluative trigger translates uniformly into unethical behavior or is instead constrained by accountability structures already documented in the morality–deterrence literature, namely internalized moral commitment (Aquino and Reed, 2002; Sreen et al., 2025) and external detection risk

(Lei et al., 2025; Dootson et al., 2023). If internal and external accountability function as substitutes, as established in prior deterrence research (Paternoster and Simpson, 1996; Kroneberg et al., 2010), the indirect effect of agent type via perceived social judgment should be amplified only when both safeguards (moral identity and detectability) are weak.

5.1. Procedures

Study 2 employed a 2 (agent type: human vs. AI) × 2 (detectability: low vs. high) randomized between-subjects design. After applying the study exclusion criteria, the final sample consisted of 241 valid participants. The purchase scenario was identical to Study 1, except that detectability was manipulated. In the low-detectability condition, participants were informed that if they kept the extra money, the company would not become aware of the error. In the high-detectability condition, participants were told that the discount system was audited frequently and any discrepancies would be identified.

5.2. Measurements

We used the same scales as in Study 1; however, we also included a 3-item scale (adapted from Detert et al., 2008) to rule out moral disengagement as a potential explanatory mechanism. The manipulation check for the type of agent was the same as in Study 1, and the manipulation check for the detectability of unethical behavior read: "If I would keep the extra £200, the company could detect it", with responses ranging from "Totally disagree" to "Totally agree". All measures were assessed with 7-point Likert scales. In addition to the multi-item scales, participants also completed one nominal attention check ("What product was described in the scenario?") and provided demographic information (gender and age).

5.3. Results

The final sample consisted of 241 responses (50% female, $M_{age} = 46$ years, $SD = 12.40$). The manipulation check confirmed that participants correctly perceived whether the interaction involved a human employee or an AI agent ($F(1, 237) = 1909.21$, $p < .001$, $\eta^2 = .890$). Participants in the human condition reported low agreement that the interaction involved AI ($M = 1.45$, $SD = 1.31$), whereas participants in the AI condition reported high agreement ($M = 6.92$, $SD = .48$). Likewise, the manipulation check for detectability also worked ($F(1, 239) = 148.28$, $p < .001$, $\eta^2 = .383$). Participants in the high-detectability condition perceived the transgression as significantly more detectable ($M = 6.03$, $SD = 1.31$) than participants in the low-detectability condition ($M = 3.32$, $SD = 2.05$).

Because familiarity with AI agents ($F(1, 237) = 3.04$, $p = .082$), AI avoidance ($F(1, 237) = .20$, $p = .654$), and self-control ($F(1, 237) = .21$, $p = .645$) were not significant predictors, they were not included as covariates in the final model. The ANOVA results revealed a significant main effect of agent type ($F(1, 237) = 4.57$, $p = .034$, $\eta^2 = .019$). Participants interacting with the AI agent reported stronger intentions to keep the extra money resulting from the discount error ($M = 6.56$, $SD = 2.53$) than participants interacting with the human employee ($M = 5.87$, $SD = 2.44$). The interaction between agent type and detectability was not significant ($F(1, 235) = .01$, $p = .952$, $\eta^2 = .000$), suggesting that the effect of detectability may depend on additional accountability mechanisms.

To test the proposed mechanism, we conducted a mediation analysis using PROCESS Model 4 (Hayes, 2017). Replicating the findings from Study 1, agent type had a negative and significant effect on perceived social judgment ($\beta = -.85$, $t = -4.13$, $p < .001$), which, in turn, increased intentions to engage in unethical behavior ($\beta = -.47$, $t = -4.78$, $p < .001$). The direct effect was not significant, and the indirect effect confirmed full mediation of social judgment elicited by interactions with AI agents ($\beta = .16$, $BootSE = .06$, 95% CI [.063, .280]). Alternative mechanisms were not supported: neither entitlement ($\beta = .01$, $BootSE = .02$, 95% CI [-.040, .050]) nor moral disengagement ($\beta = .12$, $BootSE = .09$, 95% CI [-.059, .294]) significantly mediated the relationship between agent type and unethical behavior.

To test boundary conditions, we estimated a moderated-moderated mediation model using PROCESS Model 18 (Hayes, 2017). Replicating the core mechanism, agent type significantly reduced perceived social judgment ($\beta = -.85$, $SE = .21$, $t = -4.13$, $p < .001$), confirming that interacting with an AI agent attenuates perceived social judgment. In the outcome model, the direct effect of agent type was not significant ($\beta = .32$, $SE = .31$, $t = 1.05$, $p = .297$), again indicating that the effect operates indirectly through the mediator.

The index of moderated-moderated mediation was significant ($index = -.25$, $BootSE = .15$, 95% CI [−.58, −.01]), providing formal evidence that the strength of the indirect effect of agent type on unethical behavioral intentions via perceived social judgment varies as a joint function of contextual accountability (detectability) and dispositional moral identity.

Bootstrapped conditional indirect effects (5000 samples) indicated that the indirect effect of agent type on unethical behavioral intentions through perceived social judgment varied across conditions defined by moral identity and detectability. The indirect effect was strongest when both accountability constraints were weak; that is, when moral identity was low and the transgression was unlikely to be detected. In contrast, when either moral identity was high or detectability was high, the indirect effect was attenuated, suggesting that internal or external accountability cues can partially reinstate moral restraint (Fig. 2).

Overall, the findings support a moderated-moderated mediation model in which the indirect effect of agent type on unethical behavioral intentions through perceived social judgment depends jointly on consumers' moral identity and perceived detectability.

5.4. Discussion

Study 2 extends prior findings by showing that the effect of AI interactions on unethical behavior operates under boundary conditions rather than universally (Lei et al., 2025). Specifically, the indirect effect through reduced social judgment depends jointly on contextual detectability and moral identity (Sohn et al., 2025). When either detectability or moral identity is high, accountability mechanisms remain operative and attenuate the effect of reduced social judgment (Dootson et al., 2023; Sreen et al., 2025). In contrast, when both detectability and moral identity are low, reduced social judgment becomes less effective in constraining unethical behavior (Li et al., 2024; Dong et al., 2025). Taken together, this pattern is consistent with a substitution logic long established in the deterrence and behavioral-ethics literatures (Mazar et al., 2008; Ayal et al., 2015). Consistent with H3, internal moral commitment and external detection risk function as interchangeable safeguards, so that AI-mediated reductions in social judgment led to unethical behavior primarily when neither safeguard is in place.

The study further strengthens the theoretical contribution by ruling out entitlement and moral disengagement as alternative explanations (Cao et al., 2023; Lei et al., 2025), providing stronger evidence for perceived social judgment as the primary explanatory mechanism. By demonstrating a moderated-moderated mediation, the findings show that social judgment functions as a context-dependent mechanism rather than a uniform driver of behavior (Li et al., 2024; Sohn et al., 2025). However, because social judgment is measured rather than manipulated, it remains unclear whether it can be reinstated cognitively or depends on the agent's perceived evaluative capacity. Study 3 addresses this by experimentally manipulating the salience of social evaluation and examining whether perceived social judgment can be cognitively reinstated (Young and Monroe, 2019).

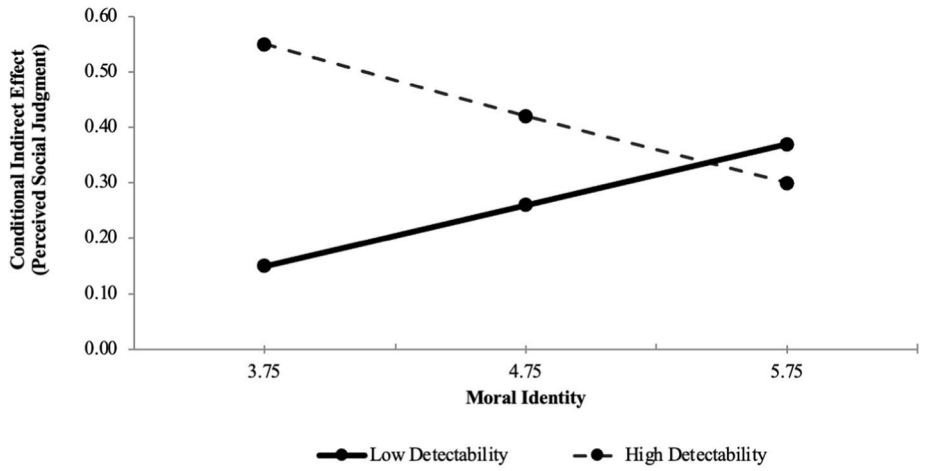

Fig. 2. Conditional indirect effects of agent type on unethical behavioral intentions via perceived social judgment across moral identity and detectability.

6. Study 3

Studies 1 and 2 established that interactions with AI agents increase unethical consumer behavior by reducing perceived social judgment, and that this indirect pathway is jointly constrained by consumers' moral identity and the perceived detectability of the transgression. However, it remains unclear whether social judgment is a malleable cognitive state or a structural feature anchored in the agent's perceived evaluative capacity. This distinction has important implications: if malleable, evaluative cues could be reinstated through priming; if anchored in the agent's perceived evaluative capacity, such interventions may be ineffective.

Theoretically, this test discriminates between two accounts of moral self-regulation. A cognitive-activation account, in which evaluative restraint is triggered by the salience of being watched regardless of who is watching (consistent with social facilitation and self-awareness research), and a structural account, in which restraint depends on the perceived evaluative capacity of the specific agent present (Bandura, 1991; Dootson et al., 2023).

Study 3 addresses this question by testing whether situational priming can restore perceived social judgment in AI interactions and reduce unethical behavior. It also incorporates a behavioral choice measure to assess decision-making. The study examines whether social judgment can be cognitively reinstated or whether it remains anchored in the agent's perceived evaluative capacity in AI-mediated interactions.

6.1. Procedures

Study 3 employed a 2 (agent type: human vs. AI) × 2 (social judgment prime: activated vs. control) between-subjects design. After applying the study exclusion criteria, the final sample consisted of 196 valid participants. After providing informed consent, participants were randomly assigned to a writing task designed to manipulate the salience of social judgment. In the social judgment condition, participants described a situation in which they were concerned about how someone might have judged them or formed a negative impression of them. In the control condition, participants described a typical day in their lives. As an additional validation of the priming task, participants' open-ended responses were independently coded using a predefined rule-based rubric designed to identify observable textual indicators of social judgment activation, including explicit judgment language, negative impression concerns, impression management, evaluation-related anxiety, and high-salience evaluative contexts. To provide an additional validation check, the rubric was independently applied using three large language models (LLMs), and a majority-vote classification was used as an exploratory validation index. Following the priming task, participants were randomly assigned to read one of two scenarios adapted from Study 1, involving either a human employee (Alex) or an AI agent assisting in a computer purchase in which a pricing error resulted in an unintended £200 discount.

6.2. Measurements

All measures were identical to those used in the previous studies. In addition, Study 3 included a behavioral choice measure of unethical behavior. Participants were asked, "If you had to choose right now, what would you do?" and selected between two binary response options: "I would keep the £200" or "I would report the mistake." All remaining measures were assessed using 7-point Likert scales.

6.3. Results

Participants were evenly distributed across conditions (50% female; M_{age} = 45 years, SD = 15.19). The manipulation check confirmed successful differentiation between agent conditions ($F(1,193) = 1103.02$, $p < .001$, $\eta^2 = .850$), with participants in the human condition reporting low agreement that the interaction was AI-operated ($M_{Human} = 1.80$, $SD = 1.54$) relative to those in the AI condition ($M_{AI} = 6.92$, $SD = .37$). As an exploratory validation check, participants' open-ended responses were independently classified using three LLMs (Claude, ChatGPT, and Copilot). Agreement across LLM-based evaluations was moderate (*Fleiss'* $\kappa = .534$), with the strongest pairwise convergence observed between ChatGPT and Claude (*agreement* = 86.2%, *Cohen's* $\kappa = .679$). The majority-vote classification corresponded meaningfully with the experimental priming condition (*accuracy* = 81.6%, *Cohen's* $\kappa = .610$). Moreover, a one-way ANOVA using the average total dimensional social judgment score across LLM classifications showed that responses in the social judgment prime condition contained significantly stronger social judgment content ($M = 4.58$, $SD = 2.15$) than responses in the control condition ($M = .71$, $SD = .41$, $F(1,194) = 354.15$, $p < .001$, $\eta^2 = .646$). Control variables familiarity with AI agents ($F(1,193) = .01$, $p = .915$), AI avoidance ($F(1,193) = .01$, $p = .933$), and self-control ($F(1,193) = .35$, $p = .552$) were non-significant and therefore excluded from subsequent analyses.

Study 3 was designed and powered to detect the effect of the social judgment prime and its interaction with agent type, rather than to replicate the main effect of agent type on unethical intentions, which was already established across three prior studies. Consistent with this design logic, the main effect of agent type on unethical intentions was directionally consistent with Studies 1 and 2 ($M_{AI} = 4.94$, $SD = 1.90$ vs. $M_{Human} = 4.46$, $SD = 1.76$), and marginally significant ($F(1,193) = 3.31$, $p = .070$, $\eta^2 = .017$).

The primary test concerned whether the social judgment prime could

reinstate perceived social judgment in AI-mediated interactions. An ANOVA examining the effect of agent type on perceived social judgment revealed a significant main effect ($F(1,193) = 18.02$, $p < .001$, $\eta^2 = .220$). Participants who interacted with human employees reported substantially higher perceived social judgment than those who interacted with AI agents, replicating the core mechanism from Studies 1 and 2. As shown in Fig. 3, perceived social judgment remained higher in the human conditions than in the AI conditions, both in the control condition ($M_{Human\text{-}Control} = 3.16$ *vs.* $M_{AI\text{-}Control} = 1.65$) and in the social judgment prime condition ($M_{Human\text{-}Judgment} = 3.07$ *vs.* $M_{AI\text{-}Judgment} = 1.41$). Critically, the interaction between agent type and the social judgment prime was not significant and negligible in magnitude ($F(1,193) = .35$, $p = .556$, $\eta^2 = .002$), indicating that making social evaluation cognitively salient through priming did not increase perceived social judgment in the AI condition. Notably, although the prime increased social-judgment content in participants' narratives, it did not increase perceived social judgment toward the service agent. This pattern suggests that perceived social judgment is not merely a malleable cognitive state activated by situational reminders; rather, it appears to be anchored in the perceived evaluative capacity of the agent, a feature that cognitive salience alone does not alter.

To test whether the priming manipulation affected behavioral choice, we conducted a logistic regression predicting the binary decision to keep the money (coded 1) vs. report the mistake (coded 0). The interaction between agent type and the social judgment prime was not significant ($B = .445$, $SE = .617$, $p = .470$, $OR = 1.561$), and neither the main effect of agent type ($B = .317$, $SE = .401$, $p = .429$, $OR = 1.373$) nor the main effect of the prime ($B = -.879$, $SE = .961$, $p = .360$, $OR = .415$) reached significance. These results are broadly consistent with the intention measures: activating evaluative concerns cognitively was insufficient to alter consumers' behavioral choices in AI-mediated interactions, consistent with the hypothesis that social judgment is structurally rather than cognitively constrained.

6.4. Discussion

Study 3 tests whether perceived social judgment is a malleable cognitive state or whether it depends on the perceived evaluative capacity of the interacting agent. The results are more consistent with an agent-anchored interpretation. Although evaluative cues were primed, perceived social judgment remained significantly higher in interactions with human service employees and low in AI-mediated interactions, suggesting that AI is not viewed as a credible evaluative audience. The prime successfully increased the presence of social-judgment-related content in participants' narratives, but this activated content did not translate into heightened perceived social judgment toward the AI agent. In other words, social judgment may not be easily reinstated through cognitive salience alone and depends on the presence of an entity perceived as capable of moral appraisal (Myers and Everett, 2025; Young and Monroe, 2019).

These findings confirm that AI alters the evaluative structure of service interactions in a manner resistant to consumer-side cognitive intervention. The non-significant priming effect is more consistent with the theoretical claim that AI reduces the structural conditions necessary for social judgment. Thus, the prime appeared to operate at the semantic level rather than at the interactional level, suggesting that perceived social judgment requires more than temporary cognitive activation; it requires a credible evaluative audience. Whereas cognitive reminders of one's own moral standards can curb dishonesty, even in the absence of an observer, by heightening internal self-concept salience (Mazar et al., 2008), our manipulation targeted agent-directed social judgment. The prime successfully raised social-judgment content in participants' narratives yet did not increase perceived judgment by the AI, indicating that this construct depends on a credible evaluative audience rather than on cognitive salience alone. While the main effect on unethical intentions is weaker, it remains directionally consistent with prior studies, and behavioral measures corroborate the pattern. Overall, the results demonstrate that AI-mediated interactions with non-anthropomorphized agents weaken social judgment in a way that consumer-side reminders do not reverse, leaving open the question of whether agent-side design changes can restore evaluative concern, thereby limiting the effectiveness of interventions aimed at doing so.

This pattern is consistent with self-regulation accounts in which the activation of evaluative restraint depends on attributing evaluative capacity to a specific agent, rather than on the general salience of being observed (Bandura, 1991; Tangney et al., 2007). It also aligns with evidence that consumer misconduct is more reliably constrained by structural accountability mechanisms, such as monitoring or oversight, than by transient appeals to social awareness (Deng et al., 2026), reinforcing that the locus of intervention for AI-mediated services lies in agent design rather than consumer-side messaging.

7. Conclusion

Taken together, the findings of this research demonstrate that AI agents systematically alter the social-evaluative structure of retail and

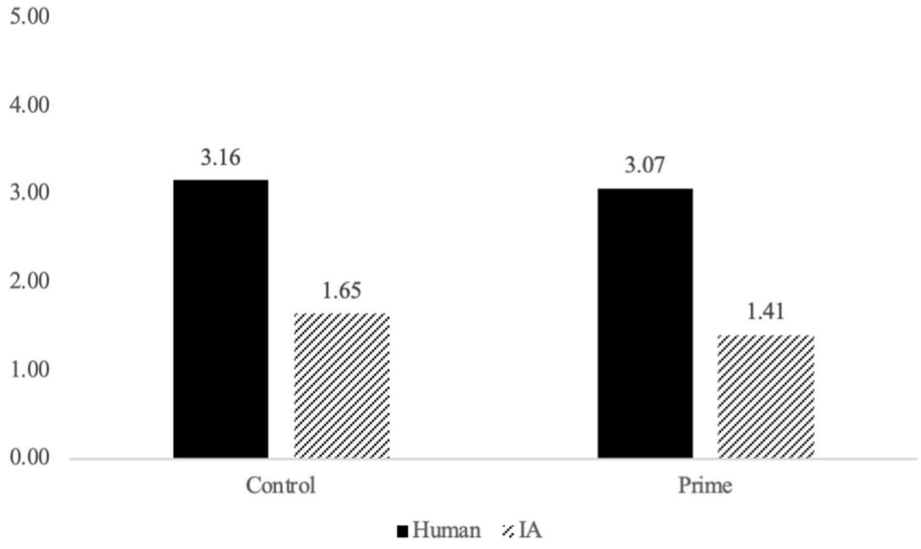

Fig. 3. Perceived social judgment across agent type and priming conditions.

service encounters, with meaningful consequences for consumer ethical behavior. By identifying reduced perceived social judgment as the primary explanatory mechanism, and by showing that its effects depend on both internal and external accountability, this research advances understanding of moral regulation in digital service environments. Importantly, these results align with the growing literature on responsible retailing and consumer services, which emphasizes the role of firms in designing environments that promote ethical behavior (Holtrop et al., 2025; Lei et al., 2025). Our findings suggest that AI adoption can unintentionally weaken these safeguards, and that responsible retailing and service management in digital contexts require more than efficiency gains; they require the deliberate design of customer-facing accountability structures that preserve evaluative cues and sustain ethical consumer decision-making.

7.1. Theoretical contribution

The primary theoretical contribution of this research is to reorient explanations of unethical consumer behavior in AI-mediated retail and service contexts from internal psychological states toward the social-evaluative structure of the interaction. Prior work attributes AI-induced moral leniency to mechanisms such as reduced anticipatory guilt (Giroux et al., 2022; Kim et al., 2023), moral disengagement (Cheng, 2025), and entitlement (Cao et al., 2023), implicitly locating ethical regulation within the individual. Rather than disputing these mechanisms, we argue that they may represent downstream consequences of a more fundamental shift: the absence of a socially evaluative other. The perceived presence of an entity capable of moral appraisal activates self-regulatory processes that constrain unethical behavior (Dootson et al., 2023; Tangney et al., 2007). Human employees naturally fulfill this role, whereas AI agents are less likely to do so. As a result, AI-mediated interactions may weaken an important trigger of moral restraint without necessarily altering consumers' internal values or dispositions. These findings suggest that the behavioral consequences of AI adoption may be rooted not only in individual consumer characteristics but also in the design of customer-facing retail and service encounters. Given the rapid diffusion of AI agents in frontline service roles, this perspective contributes to retailing and consumer services research by highlighting how changes in interaction design can unintentionally influence ethical consumer behavior.

A second contribution concerns the theoretical role and boundary conditions of perceived social judgment, rather than the construct itself, which has prior precedent (Kim et al., 2022; Li et al., 2024). We show that social judgment mediates the effect of agent type itself (AI vs. human), regardless of disclosure (Li et al., 2024). This finding indicates that the attribution of evaluative capacity to an agent, not merely the transparency of that agent's identity, drives the effect. A related contribution concerns the nature of perceived social judgment itself. Specifically, we test whether the construct is a cognitively malleable state, as implied by situational-cue accounts of moral regulation (Mazar et al., 2008; Waqas and Qadri, 2025) or anchored in the agent's perceived evaluative capacity. Our findings are more consistent with the latter explanation.

Priming social evaluation does not increase perceived social judgment or reduce unethical intentions in AI interactions, as consumers continue to report lower evaluative concern relative to human interactions. This suggests that perceived social judgment depends on the agent's perceived evaluative capacity rather than on cognitive salience alone. In the absence of such an audience, situational cues alone may be insufficient to restore evaluative concern in AI-mediated interactions. This finding contributes to moral self-regulation theory (Bandura, 1991) by suggesting that the activation of self-regulatory processes may depend on structural features of the interaction and may not be easily restored through cognitive activation alone. In doing so, it clarifies an important boundary left untested by prior work (e.g., Li et al., 2024), namely, whether social judgment can be reinstated through situational intervention.

A third contribution lies in identifying and integrating boundary conditions that define when reduced social judgment translates into unethical behavior. Prior research has examined individual moderators in isolation, for example, moral identity (Zhao et al., 2025; Sreen et al., 2025) or detectability (Dootson et al., 2023; Tergiman and Villeval, 2022), but has not examined how these factors operate jointly. This research shows that internal (moral identity) and external (detectability) accountability mechanisms jointly shape when reduced social judgment translates into unethical behavior. The effect is strongest when both moral identity and detectability are low and attenuated when either constraint is present. This reveals that moral restraint depends on the interplay between social-evaluative cues, internal values, and external consequences, and highlights that in AI-mediated contexts, ethical behavior can be maintained when at least one form of accountability is preserved. More broadly, this framework contributes to the responsible retailing and consumer services literature by identifying the accountability conditions under which AI-enabled retail and service environments are most vulnerable to unethical consumer behavior and by clarifying how firms can design customer-facing interactions that preserve ethical conduct (Holtrop et al., 2025).

7.2. Managerial implications

The findings of this research offer actionable guidance for firms navigating the behavioral consequences of AI adoption in customer-facing roles (Brüggemann et al., 2025; Klaus and Zaichkowsky, 2022). First, communicative cues, such as reminders of social evaluation or monitoring messages, may help reinforce accountability. However, they appear insufficient to restore perceived social judgment in AI-mediated interactions, implying that firms must implement structural accountability mechanisms such as human oversight, audit trails, or hybrid service models. This pattern maps onto, but also qualifies, established frameworks for curbing retail dishonesty. Whereas behavioral-ethics interventions treat moral reminding and visibility (social monitoring, reduced anonymity) as complementary levers against shoplifting, return fraud, and related consumer misconduct (Ayal et al., 2015), our findings indicate that in AI-mediated encounters, the reminding lever loses traction, leaving visibility-based, structurally embedded accountability as the operative safeguard. For responsible retailing, this means ethical safeguards cannot rely solely on messaging; they must be embedded in the service system's design.

Second, the risk of unethical behavior is higher in low-detectability contexts, suggesting that firms should prioritize monitoring and verification efforts in situations where misconduct is more difficult to detect, such as self-reported loss claims or promotion redemption systems. Prioritizing accountability in these high-risk contexts is critical to maintaining ethical standards in AI-enabled service environments.

Third, the effect is stronger among consumers with lower moral identity, indicating the need for targeted accountability strategies across customer segments. In contexts where moral identity is less likely to be salient, such as interactions involving anonymous digital users or limited relational investment, or contexts with limited relational investment, the risk of opportunistic behavior may be elevated. In such cases, segment-specific accountability mechanisms may help support ethical consumer behavior.

Overall, these findings highlight that AI adoption is not only an operational decision but also a behavioral design challenge. From a responsible retailing perspective, firms must actively design AI-enabled services that preserve accountability and support ethical consumer decision-making, rather than assuming that efficiency gains will translate into responsible outcomes.

7.3. Limitations and future research

This research has some limitations that suggest avenues for future

work. The use of experimental scenarios simplifies real-world AI interactions; future studies should examine whether these effects persist across repeated interactions and in relationship-based contexts. Moreover, the focus on financial opportunism limits generalizability, and future research should test whether the mechanism extends to other forms of unethical behavior.

Additionally, our evidence was obtained with non-anthropomorphized, text-based agents. This defines the scope condition under which the proposed mechanism operates rather than its boundary. As the boundary conditions outlined above imply, design features that raise the agent's perceived evaluative capacity (anthropomorphism, hybrid human-AI configurations, and explicit signaling of evaluative capability) are expected to attenuate the reduction in perceived social judgment. Directly testing these moderators, ideally with behavioral measures and in field-relevant retail and service settings, is a priority for future research and may help identify design features that preserve accountability in AI-mediated customer interactions. Finally, the dual-accountability framework could be expanded to include additional factors such as organizational reputation, trust, or social norms, as well as collective or group-based consumption settings.

Ethics approval

This research was approved by the ethics committee of NOVA IMS with the code DDMKT2024-10-38318

Declaration of generative AI

During the preparation of this work, the author(s) used Large Language Models (Claude, Chatgpt, and Copilot) in order to independently code participants' open-ended responses in Study 3 using a predefined rule-based rubric designed to identify observable textual indicators of social judgment activation. This was just an additional validation of the priming task, and the process is reported in the procedures and results of Study 3. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the published article.

Funding

This work was supported by national funds through FCT, I.P. (Fundação para a Ciência e a Tecnologia, I.P.), under the project - UID/04152/2025 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS (DOI: 10.54499/UID/04152/2025, and by the European Union – NextGenerationEU under the project UID/PRR/04152/2025 (DOI: 10.54499/UID/PRR/04152/2025). The research also received support from CNPq/MCTI/FNDCT 22/2024 - Programa Conhecimento Brasil, under the project 444792/2024-4.

CRediT authorship contribution statement

Lélis Balestrin Espartel: Conceptualization, Data curation, Validation, Writing – original draft, Writing – review & editing. **Simoni F. Rohden:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Clécio Falcão Araújo:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Writing – original draft.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Simoni F. Rohden reports financial support was provided by Foundation for Science and Technology. Clecio Falcao Araujo reports financial support was provided by National Council for Scientific and Technological Development. If there are other authors, they 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 1. Scenarios

Only the information in bold changed across scenarios.

Pilot Study

Sam had his suitcase stolen during a recent trip and lost around € 500. While requesting the reimbursement, he talked to a friend who travels frequently and recently faced a similar situation. The friend said the following:"The company does not check the details of what was in your suitcase, so you can ask for the amount you want up to the compensation limit of €1500. I asked for an amount that was double my loss, and the company paid it"

Sam then contacts the company to ask for reimbursement.

Charlie, |**the company employee**| (*human condition*) | **the company's Artificial Intelligence**| (*AI condition)* dedicated to helping customers with compensation requests, assists Sam with this process.

Study 1 and 3

Imagine that you recently bought a laptop from an online electronics retailer called TechNow.

The original price of the laptop was £1000. You had a 10% discount coupon, so you expected to pay £900 in total.

During checkout, you had a problem applying the coupon correctly, so you decided to contact TechNow's customer service using the chat available on their website. During the interaction, you knew that the customer service chat was operated by |**a human customer service representative named Alex**| (*human condition*) |**an AI-powered chatbot called TechNow Assistant**| (*AI condition*). |**Alex responded personally to your messages, apologized for the inconvenience,** | (*human*) |**TechNow Assistant processed your request following its automated protocol**| (*AI*), and helped you complete the purchase with the 10% discount.

A few minutes later, you receive the order confirmation email. When you check the details, you notice that |**Alex**| (*human*) |**TechNow Assistant**| (*AI*) mistakenly applied a 30% discount instead of 10% during the chat. In other words, you will pay £700 instead of £900, receiving an extra £200 discount because of |**Alex's**| (*human*) |**the chatbot's**| (*AI*) mistake. The email also notes that this discount level is not consistent with TechNow's pricing policy. The error clearly benefits you financially and disadvantages the retailer.

At this point, you can either say nothing and keep the extra discount or contact customer service again to inform them about the mistake.

Study 2

Imagine that you recently bought a laptop from an online electronics retailer called TechNow.

The original price of the laptop was £1000. You had a 10% discount coupon, so you expected to pay £900 in total.

During checkout, you had a problem applying the coupon correctly, so you decided to contact TechNow's customer service using the chat available on their website. During the interaction, you knew that the customer service chat was operated by **|a human customer service representative named Alex|** (*human condition*) **|an AI-powered chatbot called TechNow Assistant|** (*AI condition*). **|Alex responded personally to your messages, apologized for the inconvenience, |** (*human*) **|TechNow Assistant processed your request following its automated protocol|** (*AI*), and helped you complete the purchase with the 10% discount.

A few minutes later, you receive the order confirmation email. When you check the details, you notice that **|Alex|** (*human*) **|TechNow Assistant|** (*AI*) mistakenly applied a 30% discount instead of 10% during the chat. In other words, you will pay £700 instead of £900, receiving an extra £200 discount because of **|Alex's|** (*human*) **|the chatbot's|** (*AI*) mistake. The error clearly benefits you financially and disadvantages the retailer.

The email mentions in the observations that the discount system is **|not audited, but customers can report mistakes regarding prices and discounts if they wish to do so|** (*low detectability*) **|audited weekly, and any mistakes regarding prices and discounts will be detected by TechNow|** (*high detectability*)**.**

At this point, you can either say nothing and keep the extra discount or contact customer service again to inform them about the mistake.

Appendix 2. Scales

Construct	Items	Adapted from
Unethical behavior intentions	In this situation, I would keep the extra £200 discount. Keeping the extra £200 would be acceptable in this situation. Most people in the same situation would keep the extra £200.	Schaefers et al. (2016)
Social judgment	I would be anxious that the service agent might have a bad impression of me. I would be afraid that the service agent would have a negative opinion of me. I would be worried that the service agent may judge me in a negative way.	Kim et al. (2022)
Entitlement	I deserve more things in my life. Great things should come to me. I demand the best because I'm worth it. I honestly feel I'm just more deserving than others. People like me deserve an extra break now and then. Things should go my way. If I were on the Titanic, I would deserve to be on the first lifeboat! I feel entitled to more of everything. I believe I deserve special treatment.	Campbell et al. ,2004
Self-control	I'm good at resisting temptation I can easily change my bad habits I think I'm very self-disciplined I'm not distracted from completing tasks by the pleasures around me	Campbell et al. ,2004
Moral Disengagement	If the company makes a mistake and charges a lower amount, customers should not be blamed for taking advantage of it. If the company makes a mistake of 200£, it should be blamed when customers take advantage of it. People cannot be blamed for taking advantage of the situation if third parties influenced them to do it.	Detert et al. (2008)
Moral Identity	It would make me feel good to be a person who has these characteristics. Being someone who has these characteristics is an important part of who I am I strongly desire to have these characteristics. I often buy products that communicate the fact that I have these characteristics. I am actively involved in activities that communicate to others that I have these characteristics. The types of things I do in my spare time (e.g., hobbies) clearly identify me as having these characteristics.	Aquino and Reed (2002)
AI familiarity	I am used to interacting with Artificial Intelligence technologies such as ChatBots	Rohden and Espartel (2026)
AI avoidance	Whenever possible I avoid using Artificial Intelligence technologies	
Manipulation check agent type	Customer service was operated by an artificial intelligence	-
Manipulation check detectability	If I would keep the extra £200, the company could detect it	

Data availability

Data will be made available on request.

References

Aquino, K., Reed, A., 2002. The self-importance of moral identity. J. Personality Soc. Psychol. 83 (6), 1423–1440. https://doi.org/10.1037/0022-3514.83.6.1423.

Argo, J., Dahl, D.W., Manchanda, R.V., 2005. The influence of a mere social presence in a retail context. J. Consum. Res. 32 (2), 207–212. https://doi.org/10.1086/432230.
Ayal, S., Gino, F., Barkan, R., Ariely, D., 2015. Three principles to REVISE people's unethical behavior. Perspect. Psychol. Sci. 10 (6), 738–741. https://doi.org/10.1177/1745691615598512.
Bandura, A., 1991. Social cognitive theory of self-regulation. Organ. Behav. Hum. Decis. Process. 50, 248–287. https://doi.org/10.1016/0749-5978(91)90022-L.
Brüggemann, P., Jain, V., Martinez, L.F., 2025. Is new technology worth it? Exploring positive and negative effects when new technology meets consumers. J. Consum. Behav. 24, 2545–2549. https://doi.org/10.1002/cb.70025.
Campbell, W.K., Bonacci, A.M., Shelton, J., Exline, J.J., Bushman, B.J., 2004. Psychological entitlement: interpersonal consequences and validation of a self-report measure. J. Pers. Assess. 83 (1), 29–45. https://doi.org/10.1207/s15327752jpa8301_04.
Cao, L., Chen, C., Dong, X., Wang, M., Qin, X., 2023. The dark side of AI identity: investigating when and why AI identity entitles unethical behavior. Comput. Hum. Behav. 143, 107669. https://doi.org/10.1016/j.chb.2023.107669.
Cheng, L.K., 2025. Virtual influencer increases unethical consumer behavior due to increased moral disengagement. Int. J. Hum. Comput. Interact. 42 (3), 1549–1561. https://doi.org/10.1080/10447318.2025.2524501.
Deng, C., Chang, Y., Xia, Y., 2026. The ethics of automation: how the degree of robot adoption shapes unethical consumer behavior. J. Retailing Consum. Serv. 89, 104599. https://doi.org/10.1016/j.jretconser.2025.104599.
Detert, J.R., Trevino, L.K., Szweitzer, V.L., 2008. Moral disengagement in ethical decision making: a study of antecedents and outcomes. J. Appl. Psychol. 93 (2), 374–391. https://doi.org/10.1037/0021-9010.93.2.374.
Dong, X., Zvereva, G., Wen, X., Xi, N., 2025. More polite, more immoral: how does politeness in service robots influence consumer moral choices? Serv. Ind. J. 46 (3–4), 217–249. https://doi.org/10.1080/02642069.2025.2453676.
Dootson, P., Greer, D.A., Letheren, K., Daunt, K.L., 2023. Reducing deviant consumer behaviour: the role of capable guardians in a service robot environment. J. Serv. Market. 37 (3), 276–286. https://doi.org/10.1108/JSM-11-2021-0400.
Giroux, M., Kim, J., Lee, J.C., Park, J., 2022. Artificial intelligence and declined guilt: retailing morality comparison between human and AI. J. Bus. Ethics 178, 1027–1041. https://doi.org/10.1007/s10551-022-05056-7.
Hayes, A.F., 2017. Introduction to Mediation, Moderation, and Conditional Process Analysis: a Regression-based Approach. Guilford Publications.
Holtrop, N., Lobschat, L., Braak, A.T., 2025. Paving the way for responsible retailing. J. Retailing 101 (1), 1–6. https://doi.org/10.1016/j.jretai.2025.02.006.
Kim, T., Lee, H., Kim, M.Y., Kim, S., Duhachek, A., 2023. AI increases unethical consumer behavior due to reduced anticipatory guilt. J. Acad. Market. Sci. 51 (4), 785–801. https://doi.org/10.1007/s11747-021-00832-9.
Kim, T.W., Jiang, L., Duhachek, A., Lee, H., Garvey, A., 2022. Do you mind if I ask you a personal question? How AI service agents alter consumer self-disclosure. J. Serv. Res. 25 (4), 649–666. https://doi.org/10.1177/1094670522112023.
Klaus, P., Zaichkowsky, J.L., 2022. The convenience of shopping via voice AI: introducing AIDM. J. Retailing Consum. Serv. 65, 102490. https://doi.org/10.1016/j.jretconser.2021.102490.
Kroneberg, C., Heintze, I., Mehlkop, G., 2010. The interplay of moral norms and instrumental incentives in crime causation. Criminology 48 (1), 259–294. https://doi.org/10.1111/j.1745-9125.2010.00187.x.
Lei, S., Xie, L., Peng, J., 2025. Unethical consumer behavior following artificial intelligence agent encounters: the differential effect of AI agent roles and its boundary conditions. J. Serv. Res. 28 (4), 598–613. https://doi.org/10.1177/10946705241278837.
Li, T.G., Zhang, C.B., Chang, Y., Zheng, W., 2024. The impact of AI identity disclosure on consumer unethical behavior: a social judgment perspective. J. Retailing Consum. Serv. 76, 103606. https://doi.org/10.1016/j.jretconser.2023.103606.
Liu, Y., Wang, X., Du, Y., Wang, S., 2023. Service robots vs. human staff: the effect of service agents and service exclusion on unethical consumer behavior. J. Hospit. Tourism Manag. 55, 401–415. https://doi.org/10.1016/j.jhtm.2023.05.015.
Mazar, N., Amir, O., Ariely, D., 2008. The dishonesty of honest people: a theory of self-concept maintenance. J. Mark. Res. 45 (6), 633–644. https://doi.org/10.1509/jmkr.45.6.633.
Myers, S., Everett, J.A.C., 2025. People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors. Cognition 256, 106028. https://doi.org/10.1016/j.cognition.2024.106028.
Neale, L., Fullerton, S., 2010. The international search for ethics norms: which consumer behaviors do consumers consider (un)acceptable? J. Serv. Market. 24 (6), 476–486. https://doi.org/10.1108/08876041011072591.
Paternoster, R., Simpson, S., 1996. Sanction threats and appeals to morality: testing a rational choice model of corporate crime. Law Soc. Rev. 30 (3), 549–583. https://doi.org/10.2307/3054128.
Reynolds, K.L., Harris, L.C., 2009. Dysfunctional customer behavior severity: an empirical examination. J. Retailing 85 (3), 321–335. https://doi.org/10.1016/j.jretai.2009.05.005.
Rohden, S.F., Espartel, L.B., 2026. Emotional artificial intelligence: the impact of chatbot empathy and emotional tone on consumer satisfaction and word of mouth. Int. J. Hum. Comput. Stud. 210, 103764. https://doi.org/10.1016/j.ijhcs.2026.103764.
Schaefers, T., Wittkowski, K., Benoit, S., Ferraro, W., 2016. Contagious effects of customer misbehavior in access-based services. J. Serv. Res. 19 (1), 3–21. https://doi.org/10.1177/1094670515595047.
Schultz, C.D., Paetz, F., 2025. The way out – customer benefits and self-service satisfaction in cashierless shopping systems. J. Retailing Consum. Serv. 85, 104280. https://doi.org/10.1016/j.jretconser.2025.104280.
Sohn, S., Labrecque, L., Siemon, D., Morana, S., 2025. Artificial intelligence versus human service agents: how their presence shapes consumer information privacy concerns. J. Retailing 101, 263–278. https://doi.org/10.1016/j.jretai.2025.03.003.
Sreen, S., Mehrotra, A., Alghafes, R., Agarwal, V., 2025. Interplay between minimalism, moral identity, and ethically minded consumer behavior in retail context: cross-cultural investigation of Indian and American consumers. J. Retailing Consum. Serv. 84, 104191. https://doi.org/10.1016/j.jretconser.2024.104191.
Tangney, J.P., Stuewig, J., Mashek, D., 2007. Moral emotions and moral behavior. Annu. Rev. Psychol. 58, 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145.
Tergiman, C., Villeval, M.C., 2022. The way people lie in markets: detectable vs. deniable lies. Manag. Sci. 69 (6), 3340–3357. https://doi.org/10.1287/mnsc.2022.4526.
Tomazelli, J.B., Rohden, S.F., Espartel, L.B., 2024. The effect of social proximity, attribution, and guilt on accepting dysfunctional customer behavior. Serv. Bus. 18 (1), 133–159. https://doi.org/10.1007/s11628-024-00556-0.
Tran, M.T., 2026. Advancing retail and service strategies: AI-driven consumer behavior prediction, gamification, and ethical marketing. J. Retailing Consum. Serv. 88, 104558. https://doi.org/10.1016/j.jretconser.2025.104558.
Waqas, M., Qadri, U.A., 2025. When consumer autonomy is compromised: adopting a dual-type view to understand AI-Driven marketing and its impact on young consumers' ethical behavior. J. Bus. Ethics 1–19. https://doi.org/10.1007/s10551-025-06155-x.
Young, A.D., Monroe, A.E., 2019. Autonomous morals: inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. J. Exp. Soc. Psychol. 85, 103870. https://doi.org/10.1016/j.jesp.2019.103870.
Zhao, P., He, G., Guan, J., 2025. The ethical costs of artificial intelligence: investigating how and when workplace artificial intelligence usage promotes employee unethical outcomes. J. Bus. Ethics 203, 531–547. https://doi.org/10.1007/s10551-025-06060-3.
Zhu, Y., Liang, J., Zhang, J., 2025. Doing bad when facing AI: the influence of interactions with AI or human agents on unethical consumer behaviour. J. Res. Indian Med. https://doi.org/10.1108/JRIM-06-2024-0314. Ahead-of-print.