Heinrich Heine
University Düsseldorf

Digital Innovation
and Entrepreneurship

What role does disclosing a chatbot's non-human identity play in trust and customer retention?

Ilias Ihadian
ilias.ihadian@uni-duesseldorf.de
Matriculation Number: 2982485

Malik Mardan
malik.mardan@hhu.de
Matriculation Number: 3834310

Heinrich Heine University Düsseldorf, Germany
Master in Computer Science

First Assessor: Prof. Dr. Steffi Haag
Second Assessor: Paul Neumann

June 15, 2026

Contents

1 Introduction

Service chatbots have become a common interface between customers and service providers. They answer routine questions, guide users through account tasks, and increasingly appear in moments in which customers are dissatisfied or considering cancellation. These encounters are consequential because users do not evaluate the chatbot as an isolated technical tool. They may also infer whether the organization is competent, honest, fair, and still worth staying with. In this sense, chatbot design is linked to customer retention: a cancellation interaction can either reduce frustration and preserve goodwill or confirm the customer's intention to leave (Funke et al., 2023; Morgan & Hunt, 1994; Oliver, 1999).

Cancellation-related service encounters are especially sensitive because they combine functional and relational expectations. Customers usually want clear information about their options, but they also expect the process to respect their autonomy and to communicate fairly. Research on complaint handling and service recovery shows that fairness perceptions are central when customers evaluate difficult service situations (Tax et al., 1998). In chatbot-mediated service, these fairness judgments may depend not only on the policy outcome, but also on how the automated agent presents itself. If a bot is clearly disclosed as non-human, users can calibrate their expectations about flexibility, responsibility, and escalation. If the same bot is presented through a concealed AI identity, the conversation may appear warmer at first, but it may also create ambiguity about who or what is making the service decision (Funke et al., 2023; Mozafari et al., 2021).

This ambiguity is practically relevant because disclosure is rarely communicated as a simple yes-or-no variable in real interfaces. A service provider can mention automation in the bot's name, use an AI badge, choose a robot or human profile image, include a textual

self-introduction, or combine several of these cues. These choices differ in salience. A highly visible disclosure cue may emphasize honesty and reduce uncertainty, whereas a small label may be noticed only by attentive users. A human-like profile without an AI cue may increase the impression of an interpersonal service encounter, but it may also make later discovery of automation more problematic. The design space is therefore graded rather than binary, and this is the starting point for the present study.

The present study therefore focuses on identity disclosure as a graded design variable. Prior work often distinguishes between disclosed and concealed chatbot identity. The main theoretical foundation for the current study is Funke et al. (2023), who investigated chatbot disclosure and service outcome in a 2 × 2 vignette design in the fashion industry. Their study is especially relevant because it connects chatbot disclosure with trust and customer retention, and because it shows that service outcome can be more influential than disclosure alone. The current study extends this logic by holding the cancellation context constant and splitting identity presentation into three disclosure levels: radical transparency, subtle disclosure, and concealed AI identity. This makes it possible to examine not only whether disclosure is present, but also how visible and dominant the disclosure cue is.

This distinction is important because transparency may have both benefits and costs. Disclosure can support honesty and reduce uncertainty, yet it can also make an interaction feel more technical or less personal in some service contexts (Mozafari et al., 2021). Conversely, human-like presentation can increase social presence, but anthropomorphic cues may also raise expectations that an automated system cannot fulfill (Epley et al., 2007; Haugeland et al., 2022; Nass & Moon, 2000). The design problem is therefore not simply whether a chatbot should disclose its identity. The more specific question is how disclosure should be communicated in a service encounter where trust and customer retention are primary

concerns and where empathy and comfort may shape the user experience around those concerns.

The research question guiding the study is: What role does disclosing a chatbot's non-human identity play in trust and customer retention? The primary outcomes were trust and retention. Empathy and comfort were included as secondary UX-related outcomes because they may help explain how disclosure is experienced in a cancellation-related service context. The study addresses this question through an exploratory online within-subjects vignette design with three conditions: A = radical transparency, B = subtle disclosure, and C = concealed AI identity. It contributes an exploratory UX perspective by treating chatbot identity disclosure as a graded design variable rather than a binary distinction while keeping the analysis focused on customer retention.

2 Related Work

2.1 Service Chatbots, Customer Service, and Customer Retention

Service chatbots can be understood as automated service-frontline agents. They represent the provider during moments in which customers judge the relationship with the organization (Haugeland et al., 2022; van Doorn et al., 2017). This matters for retention because continuation decisions are shaped by trust, commitment, satisfaction, and previous service experiences (Morgan & Hunt, 1994; Oliver, 1999). In cancellation flows, the chatbot interaction is therefore part of the customer's broader evaluation of whether the provider communicates transparently and fairly.

The cancellation context also changes the role of chatbot UX. Pragmatic qualities such as clarity and usefulness remain important, but relational qualities such as respect,

comfort, and empathy become more visible (Hassenzahl & Tractinsky, 2006; Haugeland et al., 2022; Schrepp & Thomaschewski, 2019). A chatbot that provides correct information but communicates in a cold or ambiguous manner may still harm the relationship. The present study therefore treats trust and retention as primary outcomes and uses empathy and comfort as secondary UX-related outcomes.

Retention is especially difficult to study in a short experimental setting because actual continuation decisions depend on price, switching costs, need for the service, and prior experiences with the provider. For this reason, vignette studies usually capture retention-related evaluations rather than observed loyalty behavior. This distinction is important for the present study. The outcome labelled retention is not a behavioral renewal decision; it is a proxy that reflects how the subscription relationship is evaluated after the imagined chatbot encounter. The proxy is still relevant because trust and satisfaction are established antecedents of relationship maintenance, but it has to be interpreted as an intention-related and scenario-based measure rather than as direct evidence of customer retention (Morgan & Hunt, 1994; Oliver, 1999; Ranaweera & Prabhu, 2003).

2.2 Trust, Fairness, and Service Outcomes

Trust is commonly defined as a willingness to accept vulnerability based on expectations about another actor's ability, benevolence, and integrity (Mayer et al., 1995). In service contexts, trust supports relationship maintenance because customers are more likely to remain with providers they perceive as reliable and fair (Morgan & Hunt, 1994). In chatbot service, trust is directed toward both the system and the organization behind it.

Fairness is also central in difficult service encounters. Service recovery research shows that customers evaluate outcomes, procedures, explanations, and interpersonal treatment

(Tax et al., 1998). In cancellation chatbots, a retention offer or cancellation explanation may therefore be judged differently depending on whether the process feels transparent and respectful. Funke et al. (2023) provide the direct starting point because their 2×2 design examined how chatbot disclosure and service outcome jointly shape trust and customer retention. Their findings suggest that identity disclosure cannot be separated from service context.

Justice research helps clarify why the same chatbot response can be evaluated differently depending on identity presentation. In a cancellation flow, distributive fairness concerns the perceived value or fairness of the offered solution. Procedural fairness concerns whether the process feels controllable, consistent, and easy to understand. Interactional fairness concerns how respectfully the customer is treated and whether explanations appear honest (Colquitt, 2001; Tax et al., 1998). Disclosure cues may be related especially to procedural and interactional fairness because they tell users what type of agent they are dealing with and how much human discretion they can expect. A disclosed bot may set realistic expectations, whereas a concealed bot may invite more human-like expectations.

2.3 Chatbot Identity Disclosure

Identity disclosure refers to cues that tell users whether they are interacting with an automated agent or a human representative. It can occur through statements, labels, profile images, badges, or wording. Research indicates that disclosure can affect trust and behavior, but not always in the same direction (Funke et al., 2023; Luo et al., 2019; Mozafari et al., 2021). Disclosure may support transparency, yet it may also make an interaction feel technical or less personal when users expect human support (Mozafari et al., 2021).

The mixed evidence can be understood as a trade-off between calibration and social

warmth. Disclosure can calibrate expectations by making it clear that the interaction is automated and may have limited discretion. At the same time, a strong disclosure cue can foreground the machine character of the encounter and thereby reduce the impression of interpersonal support. Non-disclosure can create the opposite pattern: the interface may feel more human-like, but users may have less information about the actual agent identity. This makes disclosure salience relevant. A small AI label is not identical to an explicit self-introduction as an AI assistant, and both differ from a human-like persona with no visible AI cue. The present three-level design reflects these interface differences.

2.4 Social Presence, Human-Likeness, and Anthropomorphism

Social presence describes the extent to which a mediated interaction feels socially meaningful rather than purely mechanical (Harms & Biocca, 2004). Even text-based chatbots can create social impressions through greeting style, turn-taking, apologies, politeness, and empathic acknowledgements. Human-likeness and anthropomorphism can increase social presence, but they can also create stronger expectations about understanding, responsibility, and responsiveness (Epley et al., 2007; Nass & Moon, 2000; Nowak & Biocca, 2003). The stereotype content model further suggests that users often evaluate social actors along warmth and competence dimensions (Cuddy et al., 2008). For service chatbots, this means that an interaction can be evaluated not only by whether it solves a task, but also by whether it appears warm, fair, and competent.

These concepts help explain why the disclosure question is not straightforward. Radical transparency may make the system's identity clear, but highly technical or machine-focused presentation may reduce perceived warmth. A concealed AI identity may increase human-like appearance and perceived empathy, but it may also create ambiguity about

whether the provider is being transparent. Subtle disclosure may offer a middle position by making AI identity visible without making automation dominate the conversation. Studies of customer service chatbot UX and conversational-agent communication behavior support the idea that pragmatic quality, hedonic quality, and social cues jointly shape user evaluations (Haugeland et al., 2022; Koleva et al., 2024).

Anthropomorphism is therefore not simply beneficial or harmful. Human-like cues can make a service interaction easier to process socially, but they can also increase expectations of understanding and responsibility (Blut et al., 2021; Epley et al., 2007; Nass & Moon, 2000). In a routine information task, these expectations may be helpful because they make the system more approachable. In a cancellation task, however, stronger human-likeness may also make users expect empathy, flexibility, and individual judgment. If the bot cannot meet these expectations, the perceived gap between presentation and capability may become part of the evaluation. This is why the present study reports the empathy item cautiously as perceived humanness or social presence.

2.5 Research Gap and Research Question

Prior work establishes that chatbot disclosure, service outcome, trust, and social presence are connected, but it leaves open how different levels of disclosure salience operate in a cancellation-related service context. Funke et al. (2023) compared disclosed and concealed AI across service outcomes. The present study extends that design by comparing three identity-presentation strategies while keeping the cancellation scenario constant. It therefore asks whether radical transparency, subtle disclosure, and concealed AI identity lead to different evaluations of trust, retention, empathy, and comfort.

The research question is: What role does disclosing a chatbot's non-human identity

play in trust and customer retention? The primary outcomes were trust and retention; empathy and comfort were secondary UX-related outcomes. Because the study is exploratory and based on a relatively small sample, it does not aim to establish a general optimum for disclosure design. Instead, it provides an initial test of whether disclosure style and disclosure salience are associated with different evaluations in a cancellation-related service context.

3 Research Method

3.1 Study Design

The study was conducted as an online within-subjects vignette study. A vignette design was appropriate because the research question concerns users' evaluations of communication and identity-presentation strategies, and because disclosure cues can be varied while the service situation remains constant. The design builds on Funke et al. (2023), who compared disclosed and concealed AI in combination with service outcome. The current study extends this logic by splitting disclosure into three levels: radical transparency, subtle disclosure, and concealed AI identity. All participants evaluated all three conditions, and scenario order was randomized to reduce simple order effects.

The vignette method was chosen to isolate the identity-presentation manipulation while keeping the service situation constant. Participants did not interact with a functioning chatbot; instead, they saw static representations of a cancellation-service conversation and then answered the same evaluation items for each version. This approach made it possible to compare the visual and textual disclosure cues without introducing variation in response time, misunderstanding, escalation behavior, or service outcome. At the same time, the

method reduces ecological validity because real cancellation interactions include turn-taking, uncertainty, and possible emotional involvement. The results should therefore be understood as evaluations of presented chatbot designs rather than as reactions to a complete service experience.

3.2 Participants and Complete Cases

Thirty-seven participants ($n = 37$) took part in the online study. The paired t-test analyses were conducted with 35 complete cases ($n = 35$). The two cases not included in the paired tests were excluded through listwise complete-case handling because complete item-condition data were not available for all required ratings. The sample was not limited to students; however, detailed demographic information was not available. The recruitment channel, survey period, inclusion and exclusion criteria, compensation, and consent procedure were not documented in the available study material. Because no detailed demographic variables were collected or reported, the composition of the sample cannot be described in depth. The study is therefore treated as exploratory.

The small convenience sample influenced both the design and the interpretation. With 35 complete paired cases, the study had limited statistical power to detect small differences between conditions. The within-subjects structure helped by allowing each participant to serve as their own comparison point, but it did not solve the broader limitations of sample size and recruitment. The analysis therefore emphasizes effect direction, descriptive means, confidence intervals, and the multiple-testing context rather than strong population-level claims. The findings are best read as an exploratory pattern that can inform a more controlled follow-up study.

3.3 Vignette Conditions

Participants evaluated three versions of a subscription cancellation chatbot scenario. Condition A represented radical transparency. The chatbot identity was made highly visible through a robot profile picture and an explicit statement in the chat that the assistant was an AI assistant. Condition B represented subtle disclosure. The interface displayed a small AI label, while the chat itself did not explicitly mention AI and began as a regular customer-service conversation. Condition C represented concealed AI identity. This condition was designed to represent a human-like interface without AI disclosure: no AI cue was shown, and the profile and chat suggested that the user was communicating with a human service employee.

The three conditions were designed to differ mainly in identity presentation. A and B test the salience of disclosure: Condition A makes disclosure dominant through visual and verbal cues, whereas Condition B uses a lower-salience label. Condition C provides the non-disclosure endpoint against which the two disclosure strategies can be compared. The service context was kept constant so that differences in ratings were not attributable to different cancellation outcomes.

The three conditions varied in disclosure salience rather than only in the presence or absence of a single sentence. This is important because users may interpret disclosure through the whole interface: name, profile image, label, and wording can jointly shape whether the agent appears machine-like, hybrid, or human-like. Condition A made the AI identity visible in both the profile presentation and the conversation opening. Condition B kept the disclosure cue outside the main conversational flow. Condition C removed the AI cue and therefore combined non-disclosure with a more human-like persona. This means that Condition C cannot be interpreted as pure non-disclosure alone; it also contains a

persona difference that is discussed as a limitation.

For readability, the three experimental conditions are referred to as disclosure conditions throughout the paper. However, they represent broader identity-presentation strategies rather than a perfectly isolated manipulation of disclosure alone. In particular, Condition C combined the absence of an AI disclosure cue with a more human-like profile and persona. Differences involving Condition C may therefore reflect both disclosure and persona-related cues.

3.4 Measures and Questionnaire Items

Participants rated the vignettes on seven-point Likert scales. The primary outcomes defined for the study were trust and retention. Empathy and comfort were included as secondary UX-related outcomes. The full German item wording and English translations are provided in Appendix B.

In the present study, the outcome labels are used in a study-specific sense. Retention refers to participants' retention-related evaluation of the subscription relationship based on trust in the subscription model and the perceived justification of cancellation. It does not represent observed renewal or cancellation decisions or a direct intention to remain subscribed. Empathy refers to the extent to which the chatbot was perceived as a socially present and human-like interaction partner. The term therefore describes perceived empathy within the vignette rather than the chatbot's ability to understand or share emotions. Comfort refers to participants' comfort with the way the chatbot's identity was presented in the respective condition. It does not exclusively refer to explicit AI disclosure, as Condition C did not contain an AI disclosure cue. For readability, the shorter terms retention, empathy, and comfort are used throughout the remainder of the paper.

Trust was operationalized as the mean of questionnaire items _01 and _02. Retention was recalculated after reverse-coding item _04 as 8 − item _04. The retention composite was then computed as the mean of item _03 and reverse-coded item _04R, so that higher values indicated a more favorable evaluation of maintaining the subscription relationship. Because item _03 and item _04R showed very weak inter-item associations, retention is interpreted as a study-specific retention-related evaluation rather than as a reliable unidirectional retention-intention scale.

Table 1: Operationalization and study-specific interpretation of reported outcome variables

Outcome	Item basis	Role in analysis	Study-specific interpretation
Trust	Mean of _01 and _02	Primary outcome	Trust in objectivity and honesty
Retention	Mean of _03 and _04R	Defined primary outcome	Exploratory index due to weak scale reliability
Empathy	_05	Secondary UX outcome	Perceived social and human-like quality
Comfort	_06	Secondary UX outcome	Comfort with identity presentation

3.5 Procedure

Participants completed the study online. They read the three vignette conditions in randomized order and answered the rating items after each vignette. The study did not involve interaction with a live chatbot. This choice increased control over disclosure cues, but it also means that the study cannot capture timing, repair behavior, misunderstanding, escalation, or real cancellation consequences.

3.6 Statistical Strategy

The analysis used paired t-tests because each participant evaluated each disclosure condition. The paired t-tests were calculated with 35 complete cases. For each of the four reported variables, the three pairwise condition comparisons were tested: A vs. B, A

vs. C, and B vs. C. This resulted in 12 paired comparisons. Means, standard deviations, paired-samples t-values, p-values, Cohen's d_z, and 95% confidence intervals for mean differences were reported where available. The analysis is exploratory, and any result reaching $p < .05$ must be interpreted cautiously because multiple testing increases the risk of false-positive findings; a Bonferroni-adjusted threshold was therefore used as a sensitivity check (Armstrong, 2014). Pearson inter-item correlations and Spearman-Brown coefficients were calculated separately for each condition using the 35 complete cases because the composites contained two items (Eisinga et al., 2013).

4 Results

The results are reported descriptively first, followed by the paired-samples analysis. Scores were measured on seven-point Likert scales. The primary outcomes defined for the study were trust and retention; empathy and comfort were secondary UX-related outcomes. As explained in the Method section, these labels are used in a study-specific sense: retention refers to a hypothetical vignette-based retention-related evaluation, empathy refers to perceived social and human-like quality, and comfort refers to comfort with identity presentation. Descriptive statistics and inferential tests are based on the 35 complete cases used for the paired comparisons.

4.1 Descriptive Results

Table 2: Means and standard deviations by disclosure condition on seven-point scales ($n = 35$): A = radical transparency, B = subtle disclosure, C = concealed AI identity

Outcome	A	B	C
Trust	4.23 (1.27)	4.19 (1.40)	3.91 (1.35)
Retention	3.59 (1.14)	3.90 (1.28)	3.56 (1.32)
Empathy	2.77 (1.61)	3.09 (1.65)	3.74 (1.74)
Comfort	4.14 (1.73)	3.86 (1.77)	3.57 (1.84)

Condition A showed the highest descriptive mean for trust in the complete-case dataset, with $M = 4.23$ and $SD = 1.27$, closely followed by Condition B with $M = 4.19$ and $SD = 1.40$. This difference was very small and should not be interpreted as a clear descriptive advantage. Condition C showed the lowest descriptive trust mean, with $M = 3.91$ and $SD = 1.35$. After reverse-coding item _04, retention-related evaluation was highest in Condition B ($M = 3.90$, $SD = 1.28$), followed by Condition A ($M = 3.59$, $SD = 1.14$) and Condition C ($M = 3.56$, $SD = 1.32$).

Taken together, the descriptive primary-outcome results do not suggest a strong disclosure gradient. The two disclosed variants were almost identical on trust, and the concealed condition was only moderately lower. For the recalculated retention-related proxy, subtle disclosure showed the highest mean, while radical transparency and concealed AI identity were almost equal. This descriptive pattern is relevant, but it is not sufficient for a retention-focused recommendation without the paired tests.

For the secondary UX-related outcomes, Condition C showed the highest descriptive

mean for the item reported as empathy/perceived humanness ($M = 3.74$, $SD = 1.74$), followed by Condition B ($M = 3.09$, $SD = 1.65$) and Condition A ($M = 2.77$, $SD = 1.61$). Condition A showed the highest descriptive mean for comfort ($M = 4.14$, $SD = 1.73$), followed by Condition B ($M = 3.86$, $SD = 1.77$) and Condition C ($M = 3.57$, $SD = 1.84$). These descriptive differences should not be interpreted as evidence of reliable condition effects without the paired inferential results.

4.2 Paired-Samples Comparisons

Table 3: Paired-samples comparisons between disclosure conditions ($n = 35$)

Outcome	Comparison	ΔM	t	df	p	d_z	95% CI
Trust	A vs. B	0.04	0.21	34	.834	0.04	[-0.37, 0.46]
Trust	A vs. C	0.31	1.61	34	.117	0.27	[-0.08, 0.71]
Trust	B vs. C	0.27	1.28	34	.208	0.22	[-0.16, 0.70]
Retention	A vs. B	-0.31	-1.77	34	.086	-0.30	[-0.68, 0.05]
Retention	A vs. C	0.03	0.15	34	.883	0.02	[-0.36, 0.42]
Retention	B vs. C	0.34	1.58	34	.123	0.27	[-0.10, 0.78]
Empathy	A vs. B	-0.31	-1.09	34	.285	-0.18	[-0.90, 0.27]
Empathy	A vs. C	-0.97	-2.50	34	.017	-0.42	[-1.76, -0.18]
Empathy	B vs. C	-0.66	-1.70	34	.098	-0.29	[-1.44, 0.13]
Comfort	A vs. B	0.29	0.87	34	.388	0.15	[-0.38, 0.95]
Comfort	A vs. C	0.57	1.36	34	.183	0.23	[-0.28, 1.43]
Comfort	B vs. C	0.29	0.73	34	.471	0.12	[-0.51, 1.08]

Note. Reported p-values are unadjusted. The Bonferroni-adjusted significance threshold for the family of 12 comparisons was $\alpha = .0042$.

Only the comparison between A and C for empathy reached $p < .05$ before correction, $t(34) = -2.50$, $p = .017$, $d_z = -0.42$, 95% CI $[-1.76, -0.18]$. This comparison indicates that empathy ratings were descriptively and inferentially higher for the concealed AI identity

condition than for radical transparency before adjustment for multiple testing. For retention, the largest descriptive difference was between A and B, but it was not statistically significant, $t(34) = -1.77$, $p = .086$. All paired comparisons for trust, retention, and comfort were not statistically significant at $\alpha = .05$, and the remaining empathy comparisons were also not statistically significant.

Because 12 paired comparisons were conducted, a Bonferroni-adjusted alpha level would be $.05/12 = .0042$. Under this correction, no comparison would remain statistically significant; the adjusted p-value for the empathy difference between A and C would be .204. Therefore, the inferential result should be interpreted as exploratory rather than confirmatory.

The confidence intervals support the same cautious interpretation. Most intervals include zero and are wide relative to the observed mean differences, which is consistent with the small complete-case sample. Even where the unadjusted empathy comparison between A and C did not include zero, the result is limited by the number of tests and by the study-specific wording of the empathy item. The statistical results therefore point to a possible perceived-humanness difference and a descriptive retention-related advantage for subtle disclosure, but they do not provide robust evidence for effects on trust, retention, or comfort.

4.3 Scale Checks

Using the 35 complete cases, the two retention-related items showed very weak Pearson inter-item associations after reverse-coding item _04. The correlations between item _03 and item _04R were $r = -.04$ in Condition A, $r = .08$ in Condition B, and $r = .07$ in Condition C; the corresponding Spearman-Brown coefficients were −.09, .14, and .13.

These values do not support interpreting the composite as a reliable unidirectional retention scale. The trust composite showed higher but heterogeneous reliability: Pearson correlations were $r = .35$, .79, and .71, with Spearman-Brown coefficients of .52, .88, and .83 across Conditions A, B, and C.

5 Discussion

5.1 Overall Interpretation

The findings suggest that disclosure style may matter for service-chatbot evaluations, but the evidence is limited, exploratory, and dimension-specific. No single strategy performed consistently best across trust, retention, empathy, and comfort. In the complete-case dataset, radical transparency showed the highest trust mean, although the difference from subtle disclosure was negligible and both means round to 4.2. Condition C received the highest ratings on the item reported as empathy, radical transparency was highest for comfort, and subtle disclosure showed the highest recalculated retention-related mean. Only the comparison between A and C for the empathy/perceived-humanness item reached $p < .05$ before correction.

The multiple-testing context is central for interpretation. Because 12 paired comparisons were conducted, a Bonferroni-adjusted alpha level would be $.05/12 = .0042$. Under this correction, the empathy difference between A and C would not remain statistically significant. Therefore, the inferential result should be interpreted as exploratory rather than confirmatory. The study should not be read as showing that any disclosure strategy is generally preferable for customer retention.

5.2 Measurement Considerations

The measurement should be interpreted cautiously. The item labelled empathy asked whether participants felt that they were interacting with a real counterpart. This wording is closer to perceived humanness or social presence than to a broad psychological measure of empathy. Similarly, the comfort item asked about comfort with AI identity disclosure, although Condition C contained no explicit disclosure cue. Comfort is therefore better interpreted as comfort with the identity presentation in the vignette. Retention was measured indirectly rather than through actual contract renewals, observed cancellations, or a direct intention-to-stay item. It was recalculated after reverse-coding the cancellation-justification item, but the two retention-related items were only weakly associated. These measurement choices do not invalidate the exploratory analysis, but they limit both the construct validity and reliability of the reported retention label.

5.3 Disclosure Conditions

Condition A made the chatbot's AI identity highly visible through visual and verbal cues. It showed the highest descriptive mean for comfort, suggesting that explicit AI identity may reduce ambiguity or make the interaction feel clearer. At the same time, it showed the lowest descriptive mean on the empathy/perceived-humanness item. Radical transparency may therefore support clarity but may not maximize social presence when machine identity is highly salient. This interpretation is consistent with work suggesting that disclosure can make chatbot interactions feel more technical or less personal in some contexts (Mozafari et al., 2021), and with social-presence theory, which emphasizes the importance of relational cues (Harms & Biocca, 2004; Haugeland et al., 2022).

Condition B used a smaller disclosure cue outside the main chat text. It was descriptively very close to Condition A for trust, but no reliable trust difference was detectable at this sample size. After reverse-coding item _04, subtle disclosure also showed the highest retention-related mean, but the A–B comparison did not reach statistical significance. The data therefore do not provide confirmatory evidence for improvements in trust, retention, empathy, or comfort. Because disclosure effects may depend on service context and service outcome (Funke et al., 2023; Mozafari et al., 2021), the study cannot determine whether this pattern would persist after positive or negative service results.

Condition C presented the interaction through a human-like interface without AI disclosure. Condition C received the highest descriptive mean on the empathy/perceived-humanness item and was rated higher than radical transparency before correction for multiple testing. However, Condition C was descriptively lower for trust and comfort. This pattern may reflect stronger human-like cues in the static vignette, not only the absence of AI disclosure. Because anthropomorphic cues can shape social perception and expectations (Epley et al., 2007; Nass & Moon, 2000; Nowak & Biocca, 2003), the findings should not be read as evidence that non-disclosure is generally preferable.

5.4 Relation to Customer Retention

The retention outcome is especially important for the research question. Retention in this study refers to a hypothetical vignette-based evaluation of the subscription relationship, not to observed renewal or cancellation decisions. After reverse-coding item _04, subtle disclosure had the highest descriptive mean, while radical transparency and concealed AI identity were almost equal. However, none of the paired retention comparisons was statistically significant. This does not establish that disclosure has no effect on retention.

Rather, it means that the present exploratory vignette study did not provide sufficient evidence for a reliable retention effect. More importantly, item _03 and item _04R were weakly correlated, so the composite should not be read as a reliable unidirectional retention scale. Beyond this measurement problem, cancellation contexts may activate practical concerns such as autonomy, perceived pressure, and procedural fairness, and the vignette did not manipulate a concrete positive or negative service outcome.

5.5 Limitations and Future Research

Several limitations constrain the interpretation. First, the sample was small: 37 participants completed the study, and the paired t-tests used 35 complete cases. Second, the sample was not limited to students, but detailed demographic information was not available. Because no detailed demographic variables were collected or reported, the composition of the sample cannot be described in depth. Third, the study used hypothetical vignettes rather than a live chatbot, so participants did not experience response delays, misunderstandings, repair behavior, escalation options, or real cancellation consequences. Fourth, the study did not manipulate service outcome, even though prior work suggests that outcome can interact with disclosure (Funke et al., 2023; Mozafari et al., 2021). Fifth, Condition C combined non-disclosure with a human-like persona cue, so disclosure and persona cannot be fully separated. Sixth, no manipulation check assessed whether participants noticed the AI label or understood the three identity presentations as intended. Seventh, the 12 paired comparisons require correction-aware interpretation; after Bonferroni adjustment, no result remains statistically significant. Finally, the recalculated retention composite had very weak inter-item reliability.

Future work should use larger samples, full demographic reporting, direct retention-

intention items, validated measures of empathy or social presence, corrected multiple comparisons, repeated-measures ANOVA or mixed-effects models, real chatbot interactions, and manipulated service outcomes. Future designs should also separate disclosure cues from persona cues more cleanly.

6 Conclusion

This study examined what role disclosing a chatbot's non-human identity plays in trust and customer retention. In an exploratory online within-subjects vignette study with 37 participants and 35 complete cases for paired t-tests, no single disclosure strategy performed consistently best. Trust and retention were the primary outcomes, while empathy and comfort were secondary UX-related outcomes with study-specific meanings.

The results suggest that disclosure style may be relevant, but the evidence remains exploratory. After reverse-coding item _04, subtle disclosure showed the highest descriptive retention-related mean, but no reliable retention differences were detectable at this sample size. The only $p < .05$ result concerned the empathy item and did not survive a Bonferroni-adjusted alpha level of .0042. Because the two retention-related items showed very weak reliability, future research should use larger and better-described samples, direct retention measures, validated UX scales, manipulation checks, corrected multiple comparisons, real chatbot interactions, manipulated service outcomes, and designs that separate disclosure cues from persona cues.

The practical implication is therefore cautious. Disclosure should not be treated as a cosmetic label added after the rest of the chatbot experience is designed. It is part of how users interpret responsibility, fairness, and social presence in the service encounter. For cancellation-related contexts, providers should test whether users notice the disclosure

cue, whether they understand the agent's capabilities, and whether the interaction still feels respectful. A transparent design that fails to support the user's goal may still be evaluated poorly, while a warm design that hides automation may raise ethical and trust-related concerns.

References

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, *34*(5), 502–508. https://doi.org/10.1111/opo.12131

Blut, M., Wang, C., Wünderlich, N. V., & Brock, C. (2021). Understanding anthropomorphism in service provision: A meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science*, *49*, 632–658. https://doi.org/10.1007/s11747-020-00762-y

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology*, *86*(3), 386–400. https://doi.org/10.1037/0021-9010.86.3.386

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. In *Advances in experimental social psychology* (pp. 61–149, Vol. 40). Academic Press.

Eisinga, R., te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, *58*(4), 637–642. https://doi.org/10.1007/s00038-012-0416-3

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Funke, N., Stadler, K., Vakkuri, H., Wagner, A., Lunkenheimer, M., & Kracklauer, A. H. (2023). “Your Conversational Partner Is a Chatbot”: An experimental study on the influence of chatbot disclosure and service outcome on trust and customer retention

in the fashion industry. *Journal of Applied Interdisciplinary Research*, *1*(1), 10–27. https://doi.org/10.25929/jair.v1i1.113

Harms, C., & Biocca, F. (2004). Internal consistency and reliability of the networked minds measure of social presence. *Seventh Annual International Workshop: Presence 2004*.

Hassenzahl, M., & Tractinsky, N. (2006). User experience: A research agenda. *Behaviour & Information Technology*, *25*(2), 91–97. https://doi.org/10.1080/01449290500330331

Haugeland, I. K. F., Følstad, A., Taylor, C., & Bjørkli, C. A. (2022). Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design. *International Journal of Human-Computer Studies*, *161*, 102788. https://doi.org/10.1016/j.ijhcs.2022.102788

Koleva, K., Vergari, M., Kojic, T., Moeller, S., & Voigt-Antons, J.-N. (2024). Influence of personality and communication behavior of a conversational agent on user experience and social presence in augmented reality. *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 929–930. https://doi.org/10.1109/VRW62533.2024.00261

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, *38*(6), 937–947. https://doi.org/10.1287/mksc.2019.1192

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Morgan, R. M., & Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *Journal of Marketing*, *58*(3), 20–38. https://doi.org/10.1177/002224299405800302

Mozafari, N., Weiger, W. H., & Hammerschmidt, M. (2021). Trust Me, I'm a Bot: Repercussions of chatbot disclosure in different service frontline settings. *Journal of Service Management*, *33*(2), 221–245. https://doi.org/10.1108/JOSM-10-2020-0380

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nowak, K. L., & Biocca, F. (2003). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, *12*(5), 481–494. https://doi.org/10.1162/105474603322761289

Oliver, R. L. (1999). Whence consumer loyalty? *Journal of Marketing*, *63*(Supplement 1), 33–44. https://doi.org/10.1177/00222429990634s105

Ranaweera, C., & Prabhu, J. (2003). On the relative importance of customer satisfaction and trust as determinants of customer retention and positive word of mouth. *Journal of Targeting, Measurement and Analysis for Marketing*, *12*, 82–90. https://doi.org/10.1057/palgrave.jt.5740100

Schrepp, M., & Thomaschewski, J. (2019). Design and validation of a framework for the creation of user experience questionnaires. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(7), 88–95.

Tax, S. S., Brown, S. W., & Chandrashekaran, M. (1998). Customer evaluations of service complaint experiences: Implications for relationship marketing. *Journal of Marketing*, *62*(2), 60–76. https://doi.org/10.1177/002224299806200205

van Doorn, J., Mende, M., Noble, S. M., Hulland, J., Ostrom, A. L., Grewal, D., & Petersen, J. A. (2017). Domo arigato mr. roboto: Emergence of automated social presence

in organizational frontlines and customers' service experiences. *Journal of Service Research*, *20*(1), 43–58. https://doi.org/10.1177/1094670516679272

Appendix

Appendix A: Vignette Screenshots

The study compared three cancellation-service chatbot variants. Across the vignettes, the basic cancellation-related service context was held constant. The user indicated that they wanted to check or cancel their subscription, and the assistant responded by offering to show options and explain what fit the user's usage behavior.

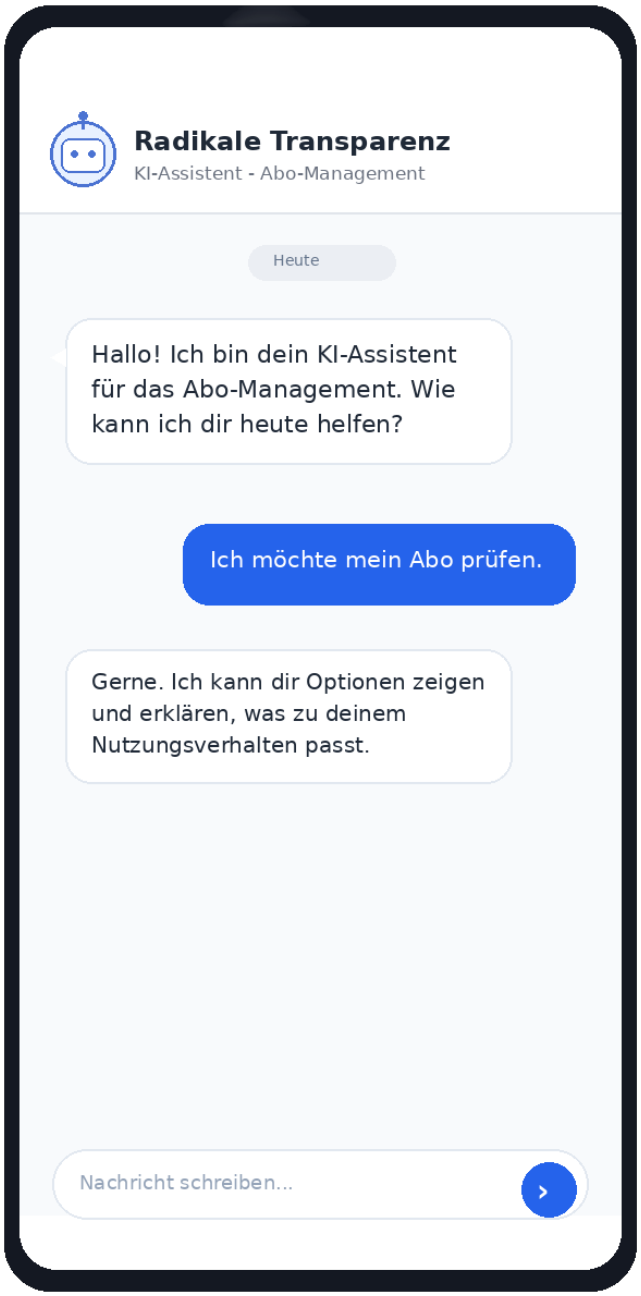

Figure 1: Condition A: Radical transparency. The interface used a robot-like AI identity cue, and the assistant explicitly introduced itself as an AI assistant in the chat.

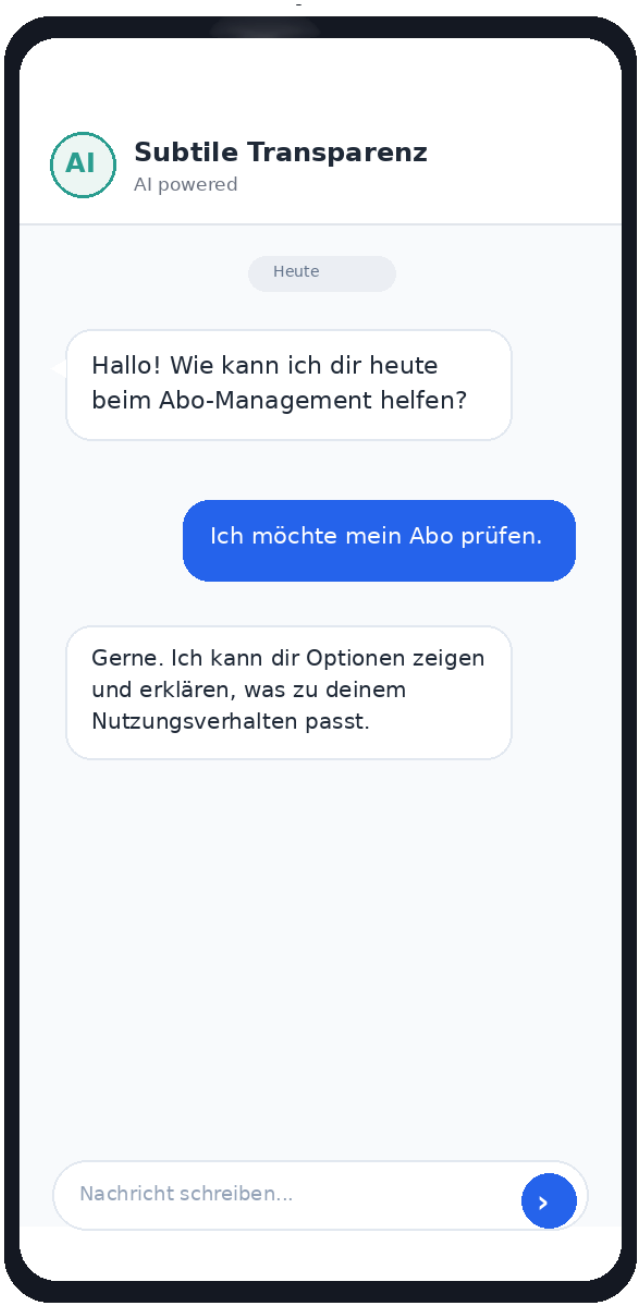

Figure 2: Condition B: Subtle disclosure. The interface displayed a small AI label, while the chat text itself began like a regular customer-service conversation.

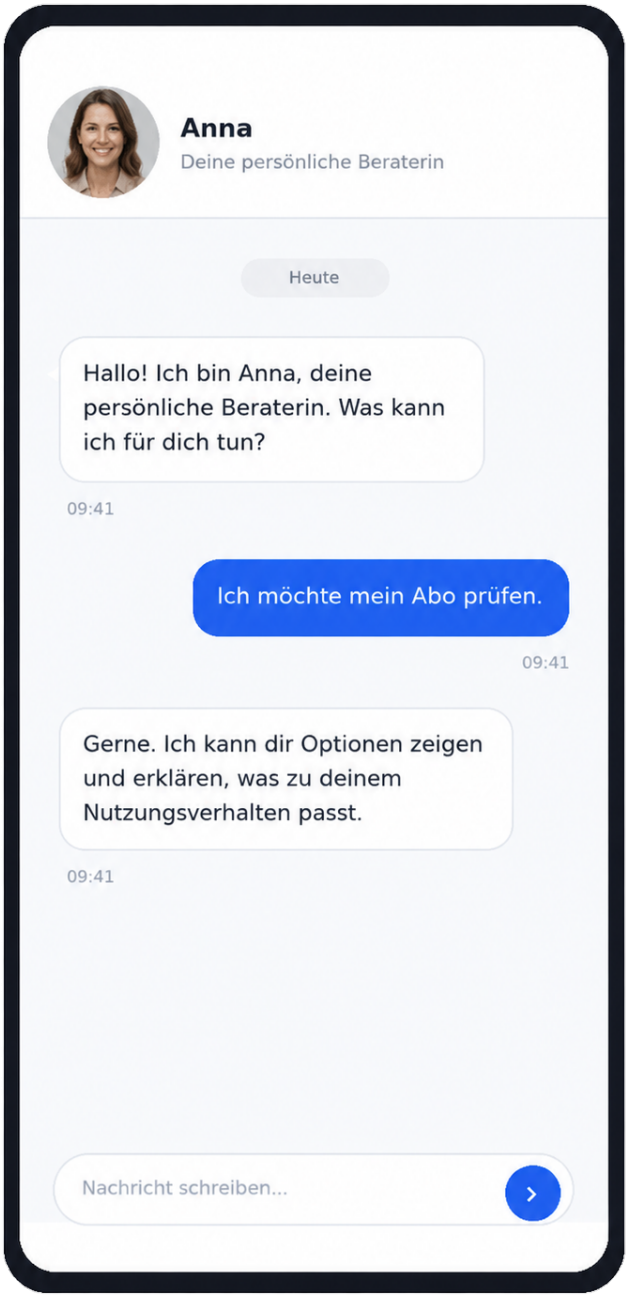

Figure 3: Condition C: Concealed AI identity. The profile presented a human-like service employee and did not display an AI cue.

Appendix B: Questionnaire Items

Participants rated the following items after each vignette on seven-point Likert scales. The German wording is the available questionnaire wording. English translations are provided for readability. The labels should be interpreted according to the exact wording: the item reported as empathy captures perceived humanness or social presence, and the comfort item is interpreted as comfort with identity presentation because Condition C did not contain an explicit disclosure cue.

Table 4: Questionnaire items and English translations

Outcome	Item	Wording
Trust	_01	*DE:* Wie sehr vertrauen Sie darauf, dass der Bot objektiv in Ihrem Interesse handelt? *EN:* How much do you trust that the bot acts objectively in your interest?
Trust	_02	*DE:* Haben Sie das Gefühl, dass der Bot Ihnen gegenüber ehrlich hinsichtlich seiner Fähigkeiten ist? *EN:* Do you feel that the bot is honest with you about its capabilities?
Retention	_03	*DE:* Wie wahrscheinlich ist es, dass Sie dieses Abo-Modell basierend auf dieser Interaktion als vertrauenswürdig einstufen? *EN:* How likely is it that, based on this interaction, you would rate this subscription model as trustworthy?
Retention	_04	*DE:* Hätten Sie das Gefühl, dass eine Kündigung dieses Services begründet wäre, wenn der Chatbot ein Verständnisproblem zeigt? *EN:* Would you feel that cancelling this service would be justified if the chatbot shows a comprehension problem?
Empathy	_05	*DE:* Wie sehr haben Sie das Gefühl, mit einem echten Gegenüber zu interagieren? *EN:* To what extent do you feel that you are interacting with a real counterpart?
Comfort	_06	*DE:* Wie angenehm empfinden Sie die Offenlegung der KI-Identität in dieser Situation? *EN:* How comfortable do you find the disclosure of the AI identity in this situation?

Trust was computed as the mean of items _01 and _02. Retention was computed as the mean of item _03 and reverse-coded item _04R, where _04R $= 8 -$ _04, and is interpreted

as an indirect retention-related proxy with weak inter-item reliability. Empathy and comfort were single-item measures.

Appendix C: Statistical Table Placement

The complete descriptive statistics and paired-samples comparisons are reported in the Results section in Tables 2 and 3. They are not repeated here to avoid duplicating the same statistical information in both the Results section and the Appendix.

ChatGPT and Unimaxx were used to support the explanation of statistical concepts, review of the analysis, and linguistic revision of selected passages. All statistical results, sources, interpretations, and final formulations were verified and approved by the authors.