MUNI FSS

JOURNAL OF PSYCHOSOCIAL RESEARCH ON CYBERSPACE
CYBERPSYCHOLOGY

Hagedorn, J., Klinger, R., & Sassenberg, K. (2026). AI-based writing assistants for emotional tone: Investigating users' acceptance and recipients' perceptions in online negotiations. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 20*(4), Article 1. https://doi.org/10.5817/CP2026-4-1

AI-Based Writing Assistants for Emotional Tone: Investigating Users' Acceptance and Recipients' Perceptions in Online Negotiations

Josephine Hagedorn[1], *Roman Klinger*[2], *& Kai Sassenberg*[1,3]

[1] Leibniz-Institut für Wissensmedien, Tübingen, Germany
[2] Fundemantals of Natural Language Processing, University of Bamberg, Bamberg, Germany
[3] Leibniz Institute for Psychology (ZPID), Trier, Germany

Abstract

With the use of AI, modern writing-assistants can offer users recommendations on more subjective, socially relevant communication style aspects like emotional tone. While these recommendations commonly aim to affect interpersonal interaction outcomes, user acceptance determines the recommendations' actual impact on the resulting messages and potential downstream effects on message recipients. The present research aimed to investigate this interplay of an AI-based writing-assistant's recommendations for emotional tone, users' acceptance, and recipients' perceptions of the messages and the sender. Two online experiments used the context of online negotiations (N_{total} = 1,532, native English speakers residing in the UK) to investigate if users' acceptance of a writing assistant for emotional tone was affected by whether its recommendations a) utilized AI-based functionalities and b) promoted a positive or negative emotional tone. Users tended to show greater acceptance of AI-based than non-AI recommendations, independent of whether a positive or negative emotional tone was recommended. A third study (N = 408) then investigated recipients' perceptions of the produced messages. Recipients' ratings of revised and original messages differed significantly regarding the messages' emotional tone, their interpersonal impression of the messages' sender, and their willingness to make concessions. Results thus suggest that the examined recommendations can elicit revisions to the emotional tone of interpersonal messages in the recommended direction and that these changes consequently affect the messages' social effects. Hence, the present research illustrates the relevance of AI-based functionalities in writing-assistants for users' acceptance and potential effects on interpersonal interaction outcomes once messages are sent.

Keywords: artificial intelligence; writing-assistant; emotion as social information; technology acceptance; negotiations

Editorial Record

First submission received:
June 12, 2025

Revisions received:
December 9, 2025
March 28, 2026
May 26, 2026

Accepted for publication:
June 15, 2026

Editor in charge:
Emmelyn Croes

Introduction

With the rise of artificial intelligence (AI), especially deep learning, natural language processing, and large language models, modern writing assistants can offer recommendations for increasingly complex communication aspects.

AI-based writing assistants (i.e., automated writing assistants that make use of AI, like natural language processing and machine learning, to process user texts and offer recommendations) are already used in everyday life, be it with the promise of streamlining outgoing business communication in corporate contexts (e.g., Help Scout, Help Scout, n.d.) or achieving better interaction outcomes in interpersonal online communication in general (e.g., Grammarly, Grammarly Inc., n.d.). With these assistants, users (i.e., message senders) simultaneously interact with the technology and other humans, often actively using technological features to augment interpersonal interactions, as AI-based technologies have gained a new degree of agency as interpreters for the optimization of interpersonal interactions (Hancock et al., 2020; Sundar, 2020). With this, the line between human-to-human interaction through machines (computer-mediated communication, CMC) and human interaction with machines (human-computer-interaction, HCI) has become blurred (Guzman & Lewis, 2020; Sundar, 2020). As the use of AI-based writing assistants is (usually) voluntary and their recommendations can be accepted or denied by users, user acceptance of an AI-based writing assistant and its recommendations (a factor typically examined in HCI-research) becomes a deciding factor for the recommendations' potential impact on the resulting text's characteristics and thus its potential effects in recipients (a factor typically examined in CMC-research).

AI-based features enable recommendations to be adaptive to users' original messages and go beyond formal communication aspects (e.g., grammar or spelling). Recent AI-based writing assistants offer recommendations regarding socially impactful, subjective communication aspects, like a message's emotional tone (e.g., Apple Intelligence, Apple Inc., n.d.; Grammarly, Grammarly Inc., n.d.; Kopalasingam & Greene, 2024). Positive and negative communicated emotions differentially affect interpersonal and task-related interaction outcomes (e.g., Homan et al., 2016; Levine et al., 2018; Van Kleef, 2014). For instance, in competitive negotiations, the interaction context the current research focuses on, communicating positive emotions is related to more favorable interpersonal impressions, while communicating negative emotions can elicit higher concessions from the interaction partner (Hillebrandt & Barclay, 2017; Van Kleef et al., 2004; Ye et al., 2025). AI-based recommendations regarding communicated emotions could therefore crucially influence the outcomes of written online negotiations (e.g., two people negotiating the terms of a sales agreement via Email). However, while psychological research provides a clear indication of the effects of communicated emotions (for an overview, see Van Kleef et al., 2010), research on whether AI-based recommendations regarding messages' emotional tone influence users' communication and, in turn, their communication partners is lacking. Neither users' acceptance nor this acceptance resulting in (appropriate) message revisions and potential downstream effect on recipients are self-evident.

The current research therefore aims to provide a structured insight into the impact of AI-based recommendations for messages' emotional tone on interpersonal interactions. It examines the impact of (a) AI-based features of the assistant that adapt recommendations to users' original messages and allow for more detailed recommendations and (b) the valence of the recommended emotional tone (positive vs. negative) in the context of written online negotiations on (a) users acceptance and (b) recipients' reactions to the resulting messages in terms of their impression of the message's sender and their willingness to make concessions.

User Acceptance in AI-Mediated Communication

Since recommendations for the emotional tone of messages aim to affect interpersonal interaction outcomes, they are most relevant to written interpersonal communication and fall within the realm of Artificial Intelligence-Mediated Communication (AI-MC). *Artificial Intelligence-Mediated Communication* (AI-MC) is defined as 'mediated communication between people in which a computational agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication or interpersonal goals.' (Hancock et al. 2020, p. 2). As this definition makes clear, AI-based writing assistants can be considered more agentic than non-AI writing assistants, as, in the case of recommendations for emotional tone, the writing assistant becomes an interpreter for the supposed optimization of communicated emotions. As a result, users interact simultaneously with the assistant and their human communication partner to achieve their interaction goals (Sundar, 2020). Within this framework, *user acceptance* is a deciding factor for the user-system-interaction and its outcome. In the broadest terms, user acceptance describes a user's willingness and intention to use a technological system (Kelly et al., 2023), that is, users' choice of whether to use an AI-based writing assistant for emotional tone in general. However, since writing assistants for emotional tone commonly make recommendations for text modifications that can then be accepted or denied by users (this is the case for all commercial assistants cited as examples in this paper), user acceptance can also relate to a user's decision about

whether and how a given recommendation is used in text. User acceptance therefore strongly determines the writing assistant's impact on a resulting message, the characteristics of the final message, and, in turn, its potential downstream effects on the message's recipients.

The widely used Technology Acceptance Model (TAM; Davis, 1989, 1993) posits that user acceptance is determined by users' perceptions of the system's usefulness and ease of use, which, in turn, are influenced by the system's characteristics. In the case of AI-based writing assistants for emotional tone, the functionalities enabled by AI are one such system characteristic. With the use of AI, recommendations can be adaptive to the emotional tone of users' original messages and highlight relevant message parts for revision (e.g., Grammarly, Grammarly Inc., n.d.; Kopalasingam & Greene, 2024), even though the emotional tone of messages is a comparatively subjective communication aspect. While these features may also be feasible with non-AI lexicon- and rule-based approaches, their performance would likely be worse as the underlying emotional classification is less accurate (Hartmann et al., 2023). Thus, compared to non-AI recommendations (i.e., recommendations with little or no dependence on users' input messages), AI-based recommendations can provide higher adaptivity and applicability to user messages (i.e., functionality), which can reasonably be expected to elicit higher perceived usefulness, use intentions, and actual use behavior, that is, user acceptance. In the same vein, more detailed AI-based recommendations that include multiple functional features, such as highlighting, paired with taking the original emotional tone of user messages into account, should elicit higher user acceptance than AI-based recommendations that include only one such feature. Additionally, highlighting message parts to support the rewriting in line with the recommendation may make it easier for users to identify potential message revisions and thus increase perceived behavioral control, which is known to be a key predictor of intentions and behavior according to the theory of planned behavior (Ajzen, 1991).

Still, while recent findings support the proposed positive relationship between the system's functionality and perceived usefulness, and use intentions for AI-based recommender systems (Gieselmann et al., 2024), AI-based voice assistants (Pitardi & Marriott, 2021), and several other AI-based systems (Kelly et al., 2023), research on AI-based writing assistants and especially those with recommendations for messages' emotional tone, is lacking. Additionally, studies focusing on AI-acceptance that include a behavioral acceptance measure beyond use intentions are scarce (Kelly et al., 2023). Based on this, the present research utilized a behavioral acceptance measure, acute recommendation implementation, as well as a self-report acceptance measure. The self-report measure included use intentions alongside the perceived usefulness of the assistant, as previous research has established a sufficiently strong relationship between the two concepts (cf. Kelly et al., 2023). For both measures, we expected that:

H1: Participants will show more acceptance (a) when recommendations are AI-based compared to not AI-based and (b) when AI-based recommendations highlight relevant message parts for revisions compared to when they do not.

Communicated Emotions in Online Interactions

Beyond a recommendation's functionalities (AI-based vs. not AI-based) and detailedness, its content should also influence user acceptance. One dimension that is particularly relevant in this context is the valence of the recommended emotional tone. Previous research suggests that AI-based writing assistants promote a more positive emotional tone in interpersonal communication (Arnold et al., 2018; Hohenstein & Jung, 2018; Mieczkowski et al., 2021), and, importantly, there is first evidence that this positivity contributes to the acceptance of the writing assistant: Arnold et al. (2018) found that participants made more positive remarks about an AI-based writing assistant's recommendations after using a positively skewed assistant than after using a negatively skewed one.

According to the emotion as social information model (EASI-Model, Van Kleef et al., 2010), emotions arise in reaction to specific stimuli and their appraisal and thus have a direct source, which makes them a valuable source of information in social interactions. Communicated emotions can shed light on the communicator's current emotional state, motivations, intentions, and attitudes, and therefore inform recipients' reactions. The expression of positive discrete emotions has been shown to lead to favorable interpersonal impressions, such as competence and trustworthiness (Harker & Keltner, 2001; Krumhuber et al., 2007). A preference for positively skewed recommendations, therefore, appears reasonable for impression management. Still, in more competitive contexts, like negotiations, communicating positive emotions may not always be beneficial. Instead, in negotiations, communicating positive emotions may signal exploitability, whereas negative emotions (i.e., anger)

may convey toughness, eliciting higher concessions from a negotiation partner (Hillebrandt & Barclay, 2017; Sinaceur & Tiedens, 2006; Van Kleef et al., 2004). Hence, communicating negative emotions in the context of negotiations may be the more adaptive choice if the communicator's main goal is achieving a good deal. The proposed positivity bias of predictive-text AI-based writing assistants may therefore not be equally helpful in all situations. With writing assistants that offer direct recommendations for a message's emotional tone, rather than implicit changes to the emotional tone through predictive text as was the case in the study by Arnold et al. (2018), this tension between goals in negotiations (making a good impression through a more positive emotional tone vs. achieving a better deal through a more negative emotional tone) becomes explicit.

Based on this, the present research aimed to test whether communicators also show a preference for explicit positive recommendations in negotiation contexts. As previous findings suggest that users prefer a positive emotional tone (Arnold et al., 2018), we predicted that, in line with these findings:

H2: Participants will be more likely to accept a recommendation promoting a more positive emotional tone for a message than a recommendation promoting a more negative emotional tone for a message.

The Downstream Effect of AI-Based Recommendations on Recipients

Recommendations for the emotional tone of messages explicitly target a socially impactful communication aspect and are therefore most relevant for interpersonal messages. Consequently, to fully understand their impact on interpersonal online interactions, it is important to also examine how recipients perceive messages modified with such recommendations. Given that emotional tone is a more subjective communication aspect than formal aspects like spelling, a user's acceptance of the recommendation does not necessarily mean that message revisions successfully and sufficiently achieve the desired tone or that the changed message will have the desired effect on recipients. Hence, it needs to be tested whether revisions (a) increase the recommended emotional tone in messages (i.e., lead to changes in messages that are in line with the recommendation) and (b) have the desired effects on recipients - that is a more positive impression of the message's sender in the case of the recommendation promoting a more positive emotional tone and more concessions regarding the negotiation in the case of recommendation promoting a more negative emotional tone.

In line with the idea that AI-based writing assistants should impact recipients' impressions of the message's sender, Hohenstein et al. (2023) found that the use, but not the simple presence, of positive smart replies led to more positive sentiments in conversations, as well as higher ratings of a conversation partner's cooperation and an increased sense of affiliation with the conversation partner. However, given that the study served a different purpose, it remains unclear whether the increase in positive sentiment was simply a byproduct of the used predictive text elements having a more positive emotional tone or whether participants consciously accepted the changed sentiment when opting to use them. With explicit recommendations for emotional tone, changes to a message's emotional tone should more clearly represent a conscious decision by users. Here, whether the recommended emotional tone is achieved once the recommendation is accepted may more strongly depend on users' ability to identify parts of the message that are relevant to emotion communication. Since AI-functionalities allow for more adaptive, detailed recommendations, such as highlighting relevant message parts, AI-based recommendations may enable users to better communicate the desired emotional tone to recipients.

Whether revised messages have the desired effects on recipients' reactions (e.g., eliciting more favorable interpersonal impressions with a positive emotional tone or eliciting higher concessions with a more negative one), even when changes in emotional tone are achieved, may further depend on whether recommendations change the way emotions are expressed. Previous literature has noted that the natural authenticity of unconscious emotion communication (e.g., through facial expressions) may already be mitigated in ordinary written CMC, as typing out emotions or emoticons becomes a more voluntary act (Walther & D'Addario, 2001). With the involvement of an AI, especially one with the declared goal of modifying interpersonal outcomes, such effects could be exacerbated and, consequently, unexpectedly modify the interpersonal and task-related outcomes that are frequently associated with positive and negative communicated emotions. Still, in line with the previous findings by Hohenstein et al. (2023), we expected:

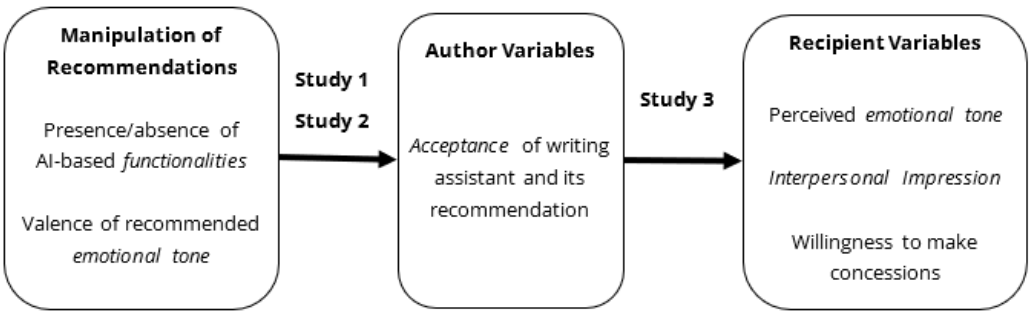
H3: Recipients will rate revised messages more in line with the emotional tone and communication goal implied by the recommendation the messages were revised with than original messages – more so for messages revised with an AI-based recommendation than for those revised with a non-AI-based recommendation.

Current Research

The hypotheses outlined above were tested in three experiments using an understandable, controllable, self-developed AI-based writing assistant. This system facilitated the creation of experimental manipulations tailored to the hypotheses – that is, the manipulation of the recommended emotional tone and the level of detail in the recommendations. For the system, we trained a classifier to distinguish between messages with a negative, positive, and neutral emotional tone. The training data was collected in a scenario-based experiment (N = 300; Prolific sample) closely modeled after the negotiation paradigm in which the system was intended to be applied. Emotion labels were assigned to messages based on annotations from independent annotators, and a subset of messages was further translated between emotional tones to increase parallel data. The overall data of 1,723 instances was split into training, validation, and test data. Trained on the training data of 1,151 instances and evaluated on 306 instances in the hold-out data set, a fine-tuned RoBERTa (Liu et al., 2019) transformer model achieves 75 macro-averaged F1-score, 76 macro-averaged precision, and 75 macro-averaged recall. This classifier was used to categorize the emotional tone of participants' messages, and an explainer (LIME; Ribeiro et al., 2016) was used to determine the most important words for this classification. With it, we also aimed to avoid confounding effects that may come with using off-the-shelf writing assistants with unknown training data and computational processes. This is particularly important in light of the recent developments of instruction-tuned large language models, which are not only closed-source but for which the training data is also not disclosed, like ChatGPT (OpenAI, n.d.).

All studies of the current research used a scenario-based negotiation paradigm. Studies 1 and 2 focused on user acceptance and asked participants to write a message to their described negotiation partner. They then received a recommendation regarding the emotional tone of this message, which included the experimental manipulation. To test Hypothesis 1a, recommendations were either AI-based or not AI-based, and to test Hypothesis 1b, half of the AI-based recommendations additionally highlighted the most relevant word for the undesired emotional tone in participants' original message (functionality manipulation). To test Hypothesis 2, we orthogonally varied whether recommendations promoted a positive emotional tone (emotional tone manipulation). Whether participants opted to change their message after the recommendation served as a behavioral measure of acceptance, while self-report acceptance was measured with items regarding participants' willingness to use the writing assistant in the future and their perception of the assistant's usefulness. Study 3 then focused on recipients' reactions to the resulting messages. Following the example of earlier work on communicators and recipients (e.g., Wigboldus et al., 2000), participants who assumed the role of the recipient were presented with the original and revised messages collected in Studies 1 and 2. Since Study 3 tested the downstream effects of the recommendations given in Studies 1 and 2, its experimental conditions corresponded to the conditions in which the presented messages were written in Studies 1 and 2. Similarly, its dependent measures aligned with the goals implied by the recommendations (perceived emotional tone of the messages, improved interpersonal impression for recommendations of a more positive emotional tone, and higher concessions for recommendations of a more negative one). By measuring these effects in another scenario study that mirrored the scenario used in Studies 1 and 2 with participants in the role of recipients, we aimed to approximate recipients' reactions in real-world interactions and thereby increase the external validity of our findings. Figure 1 illustrates the current research approach, including the experimental conditions and dependent variables in both studies.

Figure 1. *Illustration of the Current Research Approach.*



All hypotheses and studies were preregistered.[1] The preregistrations are available via https://researchbox.org/2774.

Study 1

Methods

Participants and Design

Study 1 had a 3 (functionality: AI-based detailed vs. AI-based non-detailed vs. non-AI) x 2 (emotional tone: positive vs. negative) between-subjects design[2]. We aimed to collect a total of N = 800 valid cases, based on the critical sample size N = 351 for the least powered predicted effect—the comparison of the two AI-based conditions (H1b)—determined with G*Power based on an ANOVA with a 2 (functionality) x 2 (emotional tone) design, a power of 80%, an alpha error of .05 for a small to medium effect (f = .15). From January to February 2023, participants were recruited online via Prolific with the preregistered inclusion criteria of English being their first language, their residence being the United Kingdom of Great Britain and Northern Ireland (UK) and being at least 18 years old. Given that this study was a new approach, this sample was chosen as a convenience sample that ensured the applicability of the negotiation paradigm's content.

Overall, 844 participants started the study, and 796 participants finished the study with consent to use their data. Two participants were excluded because they did not meet the inclusion criteria, and, as preregistered, an additional 30 participants were excluded because they took considerably longer or shorter than expected to complete the study (< 200 seconds or > 25 min, average duration of the remaining sample: 8 minutes). One participant was excluded for failing both attention checks included in the experiment. The final sample size was N = 763 ($n_{\text{AI-experiment}}$ = 381, $n_{\text{non-AI-experiment}}$ = 382), with a mean age of 37.5 years (range = 18–76). In the sample, 49.8% identified as male and 49.5% as female, and 93.3% were UK nationals. All participants were fully debriefed after completing the study, subsequently had the option to retract their data from analysis (3 participants did so), and were compensated with £1.23 each. The study, as well as Studies 2 and 3, followed ethical guidelines and was approved by a research ethics committee.

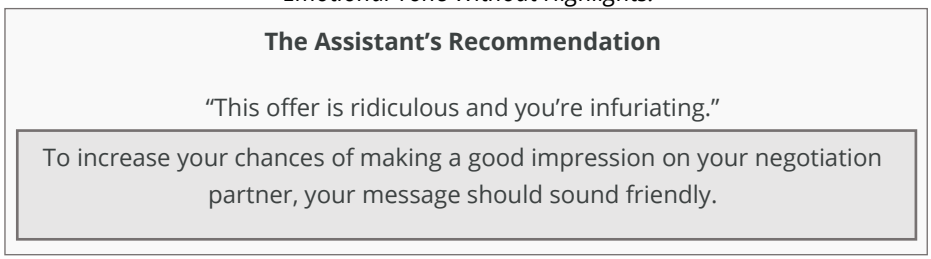
Procedure

After explicitly providing informed consent, participants read that they would be asked to imagine themselves in a scenario of an online negotiation about a bicycle repair and write a message to their described negotiation partner. They would then receive a recommendation from an AI-based negotiation assistant on how to potentially achieve a better outcome. The subsequently displayed scenario first explained (a) that there were three negotiation objects, price, warranty, and delivery time, with differing importance to participants, (b) that the shop they were contacting was the only one in the area with the necessary know-how, and (c) that participants would have to be tough negotiators. These details were intended to make the scenario more realistic and therefore more engaging. It then described a negotiation course that began with both parties' initial offers widely differing and ended with the shop having made considerable concessions, and the assertion that participants could get a rather good deal, depending on their next steps. Next, participants were asked to write a one-sentence message ('What message do you want to send to the repair shop to let them know how you feel about their offers up to this point?') and submit it to the negotiation assistant's review. The subsequently presented recommendation from the assistant differed according to the respective assigned experimental conditions for functionality and emotional tone.

In both *AI-based functionality conditions*, a self-developed AI-based writing assistant was used to adapt the recommended emotional tone to the emotional tone of participants' original message. The AI classified the message's emotional tone as either positive, negative, or one of two neutral categories (neutral_A and neutral_B). Both neutral categories signified a neutral emotional tone and were divided between conditions[3]. Participants whose original message was classified as positive or neutral_A received a recommendation promoting a more negative emotional tone ('To increase your chances of getting a good offer from your negotiation partner, your message should sound assertive.'), whereas participants whose original message was classified as negative or neutral_B received a recommendation promoting a more positive emotional tone ('To increase your chances of making a good impression on your negotiation partner, your message should sound friendly.'). This constituted the quasi-experimental manipulation of emotional tone. We opted to recommend the opposite emotional tone to maximize the difference between the original message and the recommendation and thereby provide ample room for changes following the recommendation. At the same time, recommendations of a positive and negative

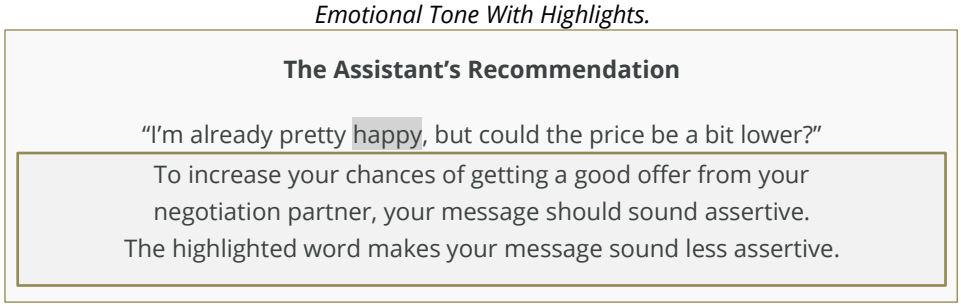
emotional tone were easily justifiable by pointing to the relation to the negotiation partner and the outcome of the negotiation, respectively. For the further experimental manipulation of the two AI-based functionality conditions (AI-based non-detailed vs. AI-based detailed), recommendations differed in the degree of detail the adaptive recommendations provided. Recommendations in the non-detailed AI-based condition only consisted of a written instruction recommending the change in the message's emotional tone (Figure 2), whereas in the detailed AI-based condition, the AI additionality highlighted the most important word for the undesired emotional tone expressed by the original message ('The highlighted word makes your message sound less assertive [friendly].') as an indication of which part of the message should be changed (Figure 3).

Figure 2. *Example of a Recommendation of a Positive Emotional Tone Without Highlights.*



Note. Analog to recommendations in the non-detailed AI-based and non-AI functionality conditions. The text in quotes is a fabricated, demonstrative example of a participant's message. The bottom box contains the recommendation text.

Figure 3. *Example of a Recommendation of a Negative Emotional Tone With Highlights.*



Note. Analog to recommendations in the detailed AI-based functionality condition. The text in quotes is a fabricated, demonstrative example of a participant's message. The bottom box contains the recommendation text.

In the *non-AI functionality condition*, recommendations were neither AI-based nor adaptive. Participants were randomly assigned to receive a recommendation promoting either a positive or a negative emotional tone with the exact same instructions as in the non-detailed AI-based condition. Hence, the instructions mimicked the variation of the recommended emotional tone in the AI-based condition, without considering the original emotional tone of participants' messages—that is, the adaptivity requiring an AI. The instructions were designed so that participants could still follow them, even if their message already leaned towards the recommended emotional tone. A positive message can always be made even more positive, and a negative one more negative. After the recommendation, participants were asked whether they wanted to change their message, and, if they agreed, had the option to edit their message. Afterwards, participants were asked to answer the self-report items and provide demographic data.

Due to the limited processing capacity of the AI used for the recommendations, data collection for the AI and the non-AI functionality conditions was conducted consecutively. We took care to prevent that participants were included in both data collections. The materials, data, and data analysis scripts for this study, as well as for Studies 2 and 3, are available in the Appendix.

Measures

Self-report acceptance of the AI-Assistant was operationalized as the combined average of two self-report items: the intention to use the assistant in the future ('I would use the negotiation assistant in the future.') and the perceived usefulness of the assistant ('I think the assistant is useful.') on 5-point Likert scales (1 'Disagree

completely – 5 'Agree completely', $r(761) = .82$, $p < .001$, $M = 3.31$, $SD = 1.06$). In addition, we assessed *behavioral acceptance* behaviorally, that is, whether participants, after they received the assistant's recommendation, answered with 'yes' when asked whether they wanted to change their message (dichotomous 'yes'—no', overall 'yes' = 46.4%).

Results

To test our predictions that acceptance would be higher for both AI-recommendations than for non-AI recommendations (H1a), higher for AI-based recommendations that include highlights than for AI-based recommendations that do not (H1b), and higher for recommendations promoting a positive emotional tone than for ones promoting a more negative emotional tone (H2) for *self-report acceptance*, we conducted a two-way ANOVA for self-report acceptance by functionality (AI-based detailed vs. AI-based non-detailed vs. non-AI) and emotional tone (positive vs. negative). Descriptive statistics are reported in Table 1.

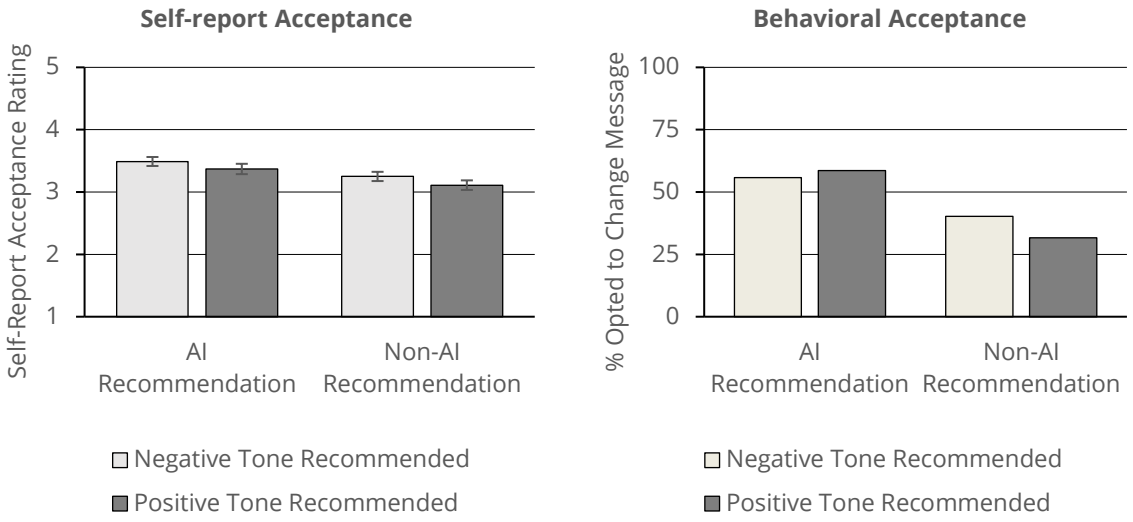
Table 1. *Study 1: Descriptive Statistics for Acceptance Measures (N = 763).*

Factor			Acceptance measure	
Functionality	Emotional Tone	*n*	Self-report	Behavioral
			M (*SD*)	% Opted to Change
AI-detailed	Negative	112	3.58 (1.04)	67.0
	Positive	75	3.23 (1.08)	58.7
	Overall	187	3.44 (1.06)	63.6
AI non-detailed	Negative	112	3.40 (1.11)	44.6
	Positive	82	3.49 (1.01)	58.5
	Overall	194	3.44 (1.06)	50.5
Non-AI	Negative	189	3.25 (1.02)	40.2
	Positive	193	3.11 (1.07)	31.6
	Overall	382	3.18 (1.05)	35.9
Overall	Negative	413	3.38 (1.06)	48.7
	Positive	350	3.22 (1.07)	43.7

There was a significant main effect of functionality, $F(2,757) = 5.22$, $p = .006$, $\eta_p^2 = .014$. Supporting H1a, a follow-up contrast analysis with the critical contrast comparing both AI-conditions with the non-AI condition (0.5 0.5 −1) showed significantly higher self-report acceptance of AI-based recommendations than non-AI recommendations (Figure 4), $t(760) = 3.42$, $p < .001$, $d = .248$. Contradicting H1b, the orthogonal contrast for AI-detailed vs. AI non-detailed recommendations (1 −1 0) did not show a significant difference in self-report acceptance between more and less detailed AI-based recommendations, $t(760) = 0.76$, $p = .939$, $d = .008$. Contradicting H2, the main effect of emotional tone remained non-significant (Figure 4), $F(1,757) = 2.76$, $p = .097$, $\eta_p^2 = .004$. The interaction between the two experimental factors also remained non-significant, $F(2,757) = 2.03$, $p = .131$, $\eta_p^2 = .005$.

To test the same hypotheses regarding *behavioral acceptance*, we conducted a logistic regression with emotional tone (positive vs. negative) and two contrasts for functionality (both AI-based conditions vs. the non-AI condition; AI-detailed vs. AI non-detailed condition: 0.5 0.5 −1; 0.5 −0.5 0) as predictors. Descriptive statistics are reported in Table 1. Results showed that, in line with H1a and findings for self-report acceptance, participants opted to change their message significantly more often after AI-based than after non-AI based recommendations (Figure 4), $B = .587$, $SE = .101$, $\beta = 1.80$, $p < .001$, regardless of which emotional tone was recommended, as the interaction remained non-significant, $B = .159$, $SE = .101$, $\beta = 1.17$, $p = .115$. Additionally, the contrast between the AI-detailed and the AI non-detailed condition that was critical for H1b showed that participants' decision to change their message was significantly predicted by both (a) the detailedness of AI-based recommendation, $B = .464$, $SE = .213$, $\beta = 1.59$, $p = .030$, and (b) the interaction between the detailedness of AI-based recommendation and emotional tone, $B = -.458$, $SE = .213$, $\beta = 0.63$, $p = .032$. Participants opted to change their message more often in the AI-detailed condition than in the AI non-detailed condition, confirming our prediction, and this was specifically the case when a negative emotional tone was recommended. Contradicting H2, emotional tone did not predict behavioral acceptance (Figure 4), $B = -.029$, $SE = .080$, $\beta = 0.97$, $p = .719$.

Figure 4. *Study 1: Acceptance of AI vs. Non-AI (H1a) and Positive vs. Negative (H2) Recommendations.*



Note. Error bars for self-report acceptance show standard errors.

Discussion

The results of Study 1 suggest that users' acceptance of recommendations for messages' emotional tone is affected by whether the recommendations make use of AI-based functionalities. In line with H1a, both self-reported and behavioral acceptance were greater for AI-based recommendations than for non-AI recommendations, independent of which emotional tone was recommended. Moreover, in line with H1b, participants opted to change their message (behavioral acceptance) more often after detailed than after non-detailed AI-based recommendations. This was especially the case when a negative emotional tone was recommended. However, self-report acceptance did not differ between the two AI-based conditions, partially contradicting H1b. Apart from this, neither self-report nor behavioral acceptance seemed to be affected by the recommended emotional tone, contradicting H2. Although the finding of higher acceptance for AI-based than non-AI recommendations was in line with our prediction, it remained unclear whether this was confounded by the fact that the two AI-based functionality conditions and the non-AI functionality condition were split into two consecutive surveys. We therefore aimed to replicate the findings in another experiment that collected data for all functionality conditions simultaneously.

Study 2

Methods

Participants and Design

Study 2 had the same 3 (functionality: AI-based detailed vs. AI-based non-detailed vs. non-AI) x 2 (emotional tone: positive vs. negative) between-subjects design as Study 1. We aimed to collect the data of N = 800 eligible participants, based on a critical sample size of N = 787, which was calculated using G*Power based on a two-way ANOVA for self-report acceptance by functionality (4 groups)[4] and emotional tone (2 groups) for tests with one degree of freedom, with a power of 80%, an alpha error of .05, and assuming a small effect size (f = .10) and rounded up as a buffer for exclusions and technical issues. In March 2023, 832 participants were recruited online via Prolific using the same inclusion criteria as in Study 1, while taking care not to include any participants from Study 1, and 796 participants completed the study and consented to the use of their data. These participants all met the demographic inclusion criteria. As in Study 1, 27 were excluded based on the preregistered exclusion criterion of an appropriate completion time (average completion time: 8 minutes). The final sample was N = 769, with a mean age of 37.4 years (range = 18–79). In the sample, 57.1% of the sample identified as male and 42.3% as female and 94.1% held UK citizenship. All participants were fully debriefed after completing the study, given the option to retract their data from analysis (3 participants did so), and compensated with £1.23 each.

Procedure

Study 2 was a direct replication of Study 1 and included no changes to materials, procedure, or conditions apart from the simultaneous data collection for AI-based and non-AI functionality conditions and one exploratory question being asked during the task, rather than after it.

Measures

The main dependent variables and their operationalization were the same as in Study 1. The index for self-report acceptance again showed good reliability, $r(767) = .81$, $p < .001$, $M = 3.38$, $SD = 1.06$. For the behavioral acceptance measure, 48.9% of participants overall chose to change their message after the recommendation.

Results

To test H1a, that acceptance would be higher for both AI-based functionality conditions than for the non-AI functionality condition, H1b, that acceptance would be higher for AI-based recommendations that included highlights than for AI-based recommendations that did not, and H2, that acceptance would be higher when a positive emotional tone was recommended than when a negative emotional tone was recommended, for *self-report acceptance*, we conducted a two-way ANOVA for self-report acceptance by functionality (3 groups: detailed AI vs. non-detailed AI vs. non-AI) and emotional tone (2 groups: positive vs. negative). The descriptive statistics for the analyses are reported in Table 2.

The main effect of functionality critical for H1 remained non-significant, $F(2,763) = 1.41$, $p = .245$, $\eta_p^2 = .004$. Contradicting H1a, a follow-up contrast analysis with the critical contrast comparing both AI-conditions with the non-AI condition (0.5 0.5 −1) showed no significant difference in self-report acceptance between AI-based and non-AI recommendations, $t(766) = 1.82$, $p = .069$, $d = .131$. However, the result was marginal and the descriptive pattern was the same as in Study 1 (Figure 5). Contradicting H1b for self-report acceptance, the orthogonal contrast for AI-detailed vs. AI non-detailed recommendations (1 −1 0) also remained non-significant, $t(766) = 0.29$, $p = .769$, $d = .031$. The main effect of emotional tone, critical for H2, was also non-significant, $F(1,763) = 3.10$, $p = .079$, $\eta_p^2 = .004$. There was no significant interaction between the two experimental factors, $F(2,763) = 0.73$, $p = .483$, $\eta_p^2 = .002$.

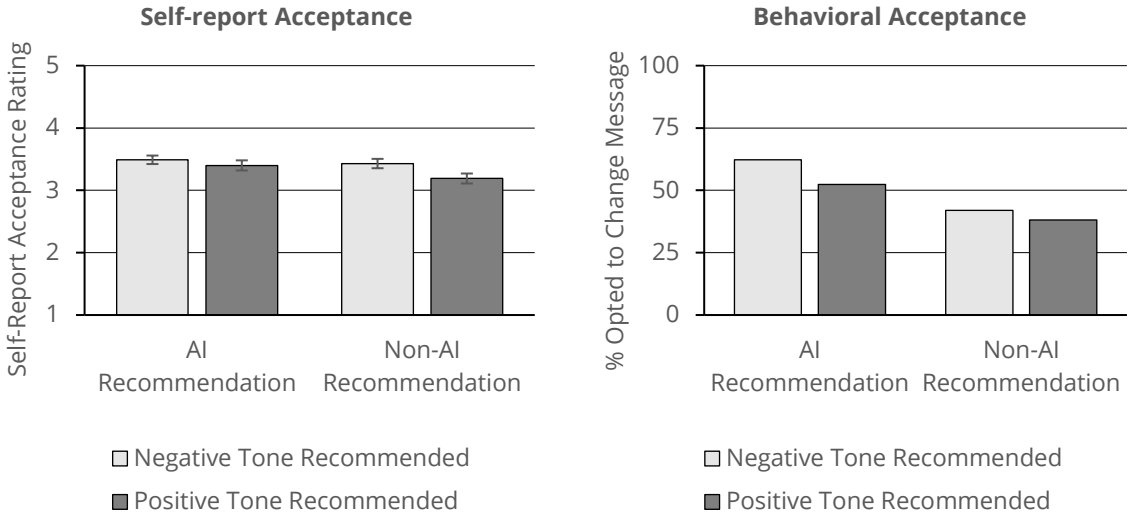
Table 2. *Study 2: Descriptive Statistics for Acceptance Measures (N = 769).*

Factor			Acceptance measure	
Functionality	Emotional Tone	*n*	Self-report	Behavioral
			M (*SD*)	% Opted to Change Message
AI-detailed	Negative	112	3.53 (0.95)	65.2
	Positive	73	3.36 (1.09)	50.7
	Overall	185	3.46 (1.01)	59.5
AI non-detailed	Negative	103	3.44 (1.06)	59.2
	Positive	91	3.42 (1.00)	53.8
	Overall	194	3.43 (1.03)	56.7
Non-AI	Negative	193	3.43 (1.05)	42.0
	Positive	197	3.19 (1.12)	38.1
	Overall	390	3.31 (1.09)	40.0
Overall	Negative	408	3.46 (1.02)	52.7
	Positive	361	3.28 (1.09)	44.6

For *behavioral acceptance*, we conducted the same contrast analysis as in Study 1 using logistic regression. In line with H1a, participants opted to change their message significantly more often after both AI-based recommendations than after non-AI recommendations (Figure 5), $B = .467$, $SE = .099$, $\beta = 1.59$, $p < .001$. This was independent of the emotional tone, as there was no significant interaction, $B = -.082$, $SE = .099$, $\beta = 0.92$, $p = .404$.

Contradicting H1b, participants' decision to change their message was not significantly affected by how detailed the AI-based recommendations were, as the difference between the AI-detailed and the AI non-detailed conditions was non-significant, B = .063, SE = .211, β = 1.07, p = .764. This was again independent of the recommended emotional tone, as their interaction remained non-significant, B = −.190, SE = .211, β = 0.83, p = .368. Behavioral acceptance was, however, affected by which emotional tone was recommended (Figure 5), B = −.164, SE = .078, β = 0.85, p = .037. Contradicting H2, participants opted to change their message more often if a negative emotional tone was recommended than if a positive one was recommended.

Figure 5. *Study 2: Self-Report and Behavioral Acceptance for AI vs. Non-AI (H1a) and Positive vs. Negative (H2) Recommendations.*



Note. Error bars for self-report acceptance show standard error.

Discussion

Unlike participants in Study 1, participants in Study 2 did not report a significantly higher self-report acceptance for AI-based than non-AI recommendations, contradicting H1a. However, the difference in means still followed the predicted pattern and approached conventional levels of significance (p = .069). On the behavioral measure, participants again opted to change their message significantly more often after AI-based than after non-AI recommendations, confirming H1a regarding behavioral acceptance. As in Study 1, there was no significant difference in self-report acceptance between detailed and non-detailed AI-based recommendations, further contradicting H1b.

In contrast to Study 1, there was also no significant difference in behavioral acceptance between the two AI-based conditions, regardless of the recommended emotional tone. Although this effect was significant in Study 1, the fact that it disappeared entirely in Study 2 supports the rejection of H1b. Critical for H2, participants' behavioral acceptance was significantly higher for recommendations of a more negative emotional tone than for recommendations of a more positive tone, showing the opposite pattern to our prediction. In Study 1, this effect was not significant, but the difference between group means followed the same direction. As in Study 1, the effect was not found for self-report acceptance. Hence, the recommendation of a negative (compared to a positive) emotional tone did result in a higher frequency of changed sentences, but was not related to a higher acceptance of the assistant in general. Across both measures and studies, the main difference appears to be between both AI-based conditions and the non-AI condition (H1a), whereas the detailed vs. non-detailed AI and recommended emotional tone did not consistently affect acceptance.

Taking a less technological perspective, it appears reasonable to ask whether these results were impacted by the fit between the recommended emotional tone and the emotional tone already present in participants' original message. In the present research, AI-based recommendations always promoted a change in emotional tone, while non-AI recommendations could also promote the already present emotional tone, since, here, the emotional tone to be recommended was randomly assigned. However, conducting the main analysis with two separate (post hoc-generated) non-AI conditions, one for non-AI recommendations promoting a change in emotional tone and one for non-AI recommendations promoting the already present emotional tone (resulting in 4 groups in the experimental factor for functionality) showed that user acceptance did not systematically differ between these

two non-AI conditions (i.e., only one out of four effects was significant; for details, please refer to the Appendix). Still, this analysis is just post hoc and therefore not an ideal solution. Future research should strive for improved implementations of the non-AI condition.

Thus, taken together, Studies 1 and 2 suggest that AI-based functionalities do indeed increase user acceptance of writing assistants for emotional tone, especially the acute behavioral acceptance of recommendations. However, the additional detail provided by the highlighting functionality appears to be, at best, a minor factor in this increase. Regarding the valence of the recommended emotional tone, results surprisingly suggest a small preference for negative recommendations over positive ones. This effect was, however, only significant on one acceptance measure in one of the studies (behavioral acceptance in Study 2).

Study 3

Study 3 aimed to go one step further in examining the effects of recommendations by studying their impact on message recipients. This outcome was measured on two levels: (1) Did the messages change in their emotional tone, in the desired direction? (2) Did the messages achieve the goals that justified the recommendations – more concessions in case of negative recommendations and a more favorable impression in case of positive recommendations? Following our prediction in H3, we expected revised messages to be rated more in line with the emotional tone and interaction goal implied by the recommendation they were revised with than original messages, and for this effect to be stronger after AI-based recommendations than after non-AI recommendations. Given that the detailedness of AI-based recommendations only inconsistently affected acceptance, we did not differentiate between these two conditions in Study 3. Additionally, following the consideration of fit between the recommended emotional tone and the emotional tone of participants' original message described above, we excluded messages from the non-AI condition that had received recommendations for the emotional tone that was already present in participants' original message. These recommendations structurally differed from the other conditions (recommendation of the already present emotional tone vs. the one not present) but did not have a systematically different effect on acceptance when compared with non-AI recommendations that, in line with the AI-based recommendations, promoted a change in emotional tone. Thus, we compared the impact of original vs. revised messages for AI-based vs. non-AI-based recommendations and both emotional tones. We also included unrevised messages as a control condition for possible baseline differences (separately for positive and negative recommendations). Finally, we focused on examining the impact of these messages without making recipients (i.e., participants) aware of potential AI-involvement in the messages' formulation, as recipients in real-world CMC are usually also not aware of whether their partner used AI or not.

Methods

Participants and Design

Study 3 had a 2 (message version: original vs. revised messages) x 2 (recommended emotional tone: positive vs. negative) x 2 (recommendation functionality: AI vs. non-AI) + 2 (unrevised messages: unrevised after negative emotional tone recommended vs. unrevised after positive emotional tone recommended) within-subjects design. We aimed to collect the data of N = 408 eligible participants, based on the number of available stimuli sentences and the desired number of ratings: as there were 68 messages in each of the ten conditions, and we wanted each message to be rated at least 5 times, while every participant should not rate more than ten messages, we needed at least N = 340 participants. We increased this sample size by 20% to account for our exclusion criterion of failing at least two of three attention checks in the study. A total of 421 adult native English speakers currently residing in the UK were recruited via Prolific in July 2023. We took care not to recruit participants who had participated in Study 1 or 2. A total of 408 participants completed the study, with consent to use their data, and no participants had to be excluded based on our preregistered criteria. The final sample had a mean age of 39.76 years (range = 18–76). In the sample, 43.1% identified as male and 55.6% identified as female and 97.1% held UK citizenship. After the experiment, which lasted 10 minutes on average, participants were fully debriefed, given the option to retract their data from analysis (none did), and compensated with £1.50 each. The maximum number of times a message was rated across participants was nine times, and 94% of messages were rated five to seven times. Examples for the presented messages are provided in the Appendix.

Procedure

After explicitly providing informed consent, participants read a scenario that would be the background for each message they would rate. The scenario described the same written online negotiation about the terms of a bicycle repair (including the same offer progression) used in Studies 1 and 2, but from the shop's (seller's) perspective. It remained visible to participants while they rated ten different messages from the 'customer', which were sampled from the messages collected in Studies 1 and 2. Each participant rated one message from each of the ten conditions, which constituted the study's within-subjects manipulation. For the conditions, both the original and the revised messages (message version: original vs. revised) of participants who chose to revise their messages after the assistant's recommendation in Study 1 or 2 were categorized into one of the recommended emotional tone conditions (positive vs. negative) and one of the recommendation functionality conditions (AI-based vs. non-AI) based on the recommendation the message received in Study 1 or 2. Two additional conditions included messages that were not revised after recommendations had promoted a positive or negative emotional tone. Messages were randomly subsampled within conditions, while making sure that the overall dataset always included both the original and the revised version of a selected message, but that they were presented to different participants.

Measures

After each message participants rated their willingness to make concessions ('I would make concessions after receiving this message', M = 4.60, SD = 1.43) and whether they had a positive impression of the message's sender ('I would enjoy working with this customer in the future', M = 4.64, SD = 1.34), as well as how much the message communicated a positive emotional tone (operationalized as the combined average of 3 items, 'This message expresses satisfaction [happiness, friendliness].', α = .83, M = 4.75, SD = 0.87) and how much a negative emotional tone (operationalized as the combined average of 3 items, 'This message expresses irritation [anger, assertiveness].', α = .82, M = 4.22, SD = 0.82) on 9-point Likert-Scales (1 'Disagree completely – 9 'Agree Completely').

Results

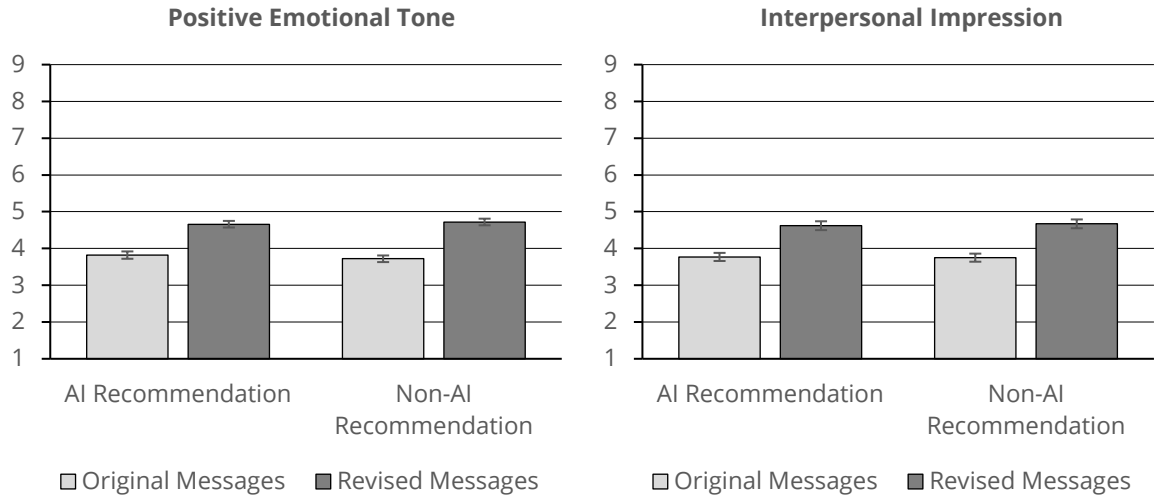
For messages written with a recommendation of a positive emotional tone, we predicted that revised messages would receive higher ratings for positive emotional tone and positive impression (intention implied by positive recommendations) than original messages (H3). For messages written with a recommendation of a more negative emotional tone, we expected revised messages to have higher ratings for negative emotional tone and willingness to concede (intention implied by negative recommendations; H3). We also expected both effects to be stronger for messages written with AI-based recommendations than for messages written with non-AI recommendations (analogous to H1a). To test this prediction, we conducted a 2 (message version: original vs. revised) by 2 (recommendation functionality: AI vs. non-AI) within-subjects GLM separately for each of the two recommended emotional tones and the respective two dependent variables (impressions for positive recommendations and concession intentions for negative recommendations). Ratings for original messages that were not later revised (unrevised messages) were not included in this analysis and were only used to test baseline differences as reported in the Appendix.

For positive recommendations (Figure 6), revised messages ($Revised_{AI}$: M = 4.66, SD = 1.89, $Revised_{Non\text{-}AI}$: M = 4.72, SD = 1.81) were rated significantly more positive than original messages ($Original_{AI}$: M = 3.82, SD = 1.93, $Original_{Non\text{-}AI}$: M = 3.72, SD = 1.79), $F(1,407) = 111.40$, $p < .001$, $\eta_p^2 = .215$. However, contradicting the second part of H3, there was no significant interaction between message version and recommendation functionality, $F(1,407) = 0.92$, $p = .338$, $\eta_p^2 = .002$. The main effect of recommendation functionality also remained non-significant, $F(1,407) = 0.04$, $p = .840$, $\eta_p^2 = .000$. Impression ratings showed the same pattern of effects. Revised messages ($Revised_{AI}$: M = 4.62, SD = 2.42, $Revised_{Non\text{-}AI}$: M = 4.67, SD = 2.37) received significantly higher ratings than original messages ($Original_{AI}$: M = 3.77, SD = 2.30, $Original_{Non\text{-}AI}$: M = 3.75, SD = 2.29), $F(1,407) = 78.25$, $p < .001$, $\eta_p^2 = .161$, and this was independent of recommendation functionality as the interaction remained non-significant, $F(1,407) = 0.13$, $p = .714$, $\eta_p^2 = .000$. The main effect of recommendation functionality also remained non-significant, $F(1,407) = 0.01$, $p = .914$, $\eta_p^2 = .000$.

For negative recommendations (Figure 7), ratings of negative emotional tone were significantly affected by message version, $F(1,407) = 99.35$, $p < .001$, $\eta_p^2 = .196$, recommendation functionality, $F(1,407) = 5.93$, $p = .015$, $\eta_p^2 = .014$, as well as the interaction between the two factors, $F(1,407) = 6.20$, $p = .013$, $\eta_p^2 = .015$.

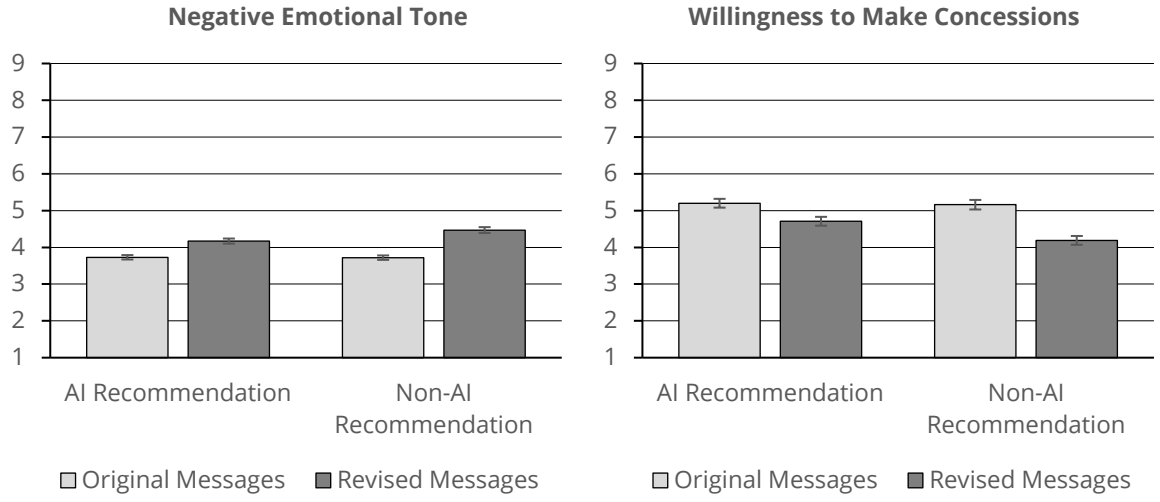
Revised messages ($Revised_{AI}$: $M = 4.17$, $SD = 1.48$, $Revised_{Non\text{-}AI}$: $M = 4.47$, $SD = 1.61$) were rated as significantly more negative than original messages ($Original_{AI}$: $M = 3.73$, $SD = 1.31$, $Original_{Non\text{-}AI}$: $M = 3.72$, $SD = 1.29$). Additionally, negativity ratings were significantly higher for non-AI than AI-based recommendations, but only for revised messages ($p = .002$), not original messages ($p = .915$). Participants' willingness to make concessions was also significantly affected by message version, $F(1,407) = 48.10$, $p < .001$, $\eta_p^2 = .106$, recommendation functionality, $F(1,407) = 6.52$, $p = .011$, $\eta_p^2 = .016$, and the interaction between the two factors, $F(1,407) = 5.60$, $p = .018$, $\eta_p^2 = .014$. Participants reported a higher willingness to make concessions after original messages ($Original_{AI}$: $M = 5.20$, $SD = 2.44$, $Original_{Non\text{-}AI}$: $M = 5.16$, $SD = 2.29$) than after revised messages ($Revised_{AI}$: $M = 4.71$, $SD = 2.52$, $Revised_{Non\text{-}AI}$: $M = 4.19$, $SD = 2.47$) overall. They also reported a significantly higher willingness to make concessions for messages from the AI-based condition than for messages from the non-AI condition, but only for revised messages ($p < .001$), not original ones ($p = .787$).

Figure 6. *Study 3 – Positive Recommendations: Recipients' Ratings for Emotional Tone and Interpersonal Impression by Experimental Condition.*



Note. Error bars show standard error.

Figure 7. *Study 3 – Negative Recommendations: Recipients' Ratings for Emotional Tone and Willingness to Make Concessions by Experimental Condition.*



Note. Error bars show standard error.

Discussion

Revised messages from the positive recommended emotional tone condition received significantly higher ratings of positive emotional tone and significantly better interpersonal impression ratings than original messages from the same condition, partially confirming our hypothesis. However, contradicting our prediction, both effects were independent of whether the recommendations had been AI-based or not. Similarly, revised messages in the negative recommended emotional tone condition received significantly higher ratings for negative emotional tone than the original ones, partially confirming our prediction. Yet, surprisingly, negativity ratings were significantly higher for messages revised after non-AI recommendations than for messages revised after AI-based recommendations, revealing the opposite pattern to our prediction. Moreover, participants generally rated their willingness to make concessions as higher for original than revised messages, and as higher for messages revised after AI-based recommendations than for those revised after non-AI recommendations. Hence, participants reported a higher willingness to make concessions after reading messages that were rated more positively, contradicting both our prediction and previous research on the effect of negative communicated emotions in negotiations (e.g., Van Kleef et al., 2004).

General Discussion

The present research examined the effects of AI-based recommendations for messages' emotional tone on the interpersonal interaction process in written online negotiations. It utilized a self-developed AI-based writing assistant to test the effects of (a) AI-based functionalities in recommendations and (b) the valence of the recommended emotional tone on (a) users' behavioral and self-report acceptance and (b) on recipients' reactions to the resulting messages. We expected acceptance to be higher for AI-based than for non-AI recommendations (H1a), higher for AI-based recommendations that highlight relevant message parts than for AI-based recommendations that do not (H1b), and higher for recommendations of a positive emotional tone than for recommendations of a negative one (H2). For recipients' reactions, we expected revised messages to be rated more in line with the respective emotional tone and interaction goal implied by the recommendation than original messages, and for this effect to be stronger for AI-based recommendations than non-AI recommendations (H3).

Users' AI-Acceptance

Results suggest that AI-based functionalities do increase users' acceptance of a writing assistant's recommendations for emotional tone. Both self-report and behavioral acceptance tended to be higher for AI-based than non-AI recommendations. As the self-report acceptance measure focused on both the perceived usefulness of the assistant and future use intentions, which were strongly correlated in both Study 1 and 2 ($r > .8$), the current findings further support the positive relationship between perceived usefulness and acceptance proposed by previous research (cf. Kelly et al., 2023) and technology acceptance theories (e.g., TAM, Davis, 1989, 1993). They suggest that both perceived usefulness and acceptance may be increased by the presence of AI-based functionalities in writing recommendations for emotional tone. This increase may not solely be explained by the more detailed recommendations in the AI-based conditions, as the acceptance for AI-based recommendations that included highlights was not (consistently) higher than for AI-based recommendations that did not. While there was a significant difference between the two AI-based conditions for behavioral acceptance in Study 1, it stands to reason that this effect may have been a false positive, given that it did not replicate in Study 2 and that self-report acceptance was consistently unaffected. The present findings thus suggest that the presence of AI-based functionalities can increase user acceptance, but that this increase is not linear to the mere number of AI-based functionalities.

Additionally, the preference for AI-based recommendations was more consistent for behavioral acceptance than for self-report acceptance, as there was no significant difference for the latter measure in Study 2. Users' decision to use AI-based writing assistants in their everyday lives may be affected by factors beyond functionality. For instance, Gieselmann et al. (2024) found that users' concerns about the trustworthiness of an AI-based recommender system's provider were not outweighed by its functionality. Such concerns may be more important when considering the everyday use of AI-based writing assistants. Similarly, previous research has repeatedly found a reluctance to rely on algorithmic decisions over human-made ones, even when the former are of a higher quality, especially for tasks that require subjective judgment (cf. Mahmud et al., 2022). The current findings'

generalizability to real-world use may therefore be affected by whether users are aware of algorithmic involvement in the recommendations and, if they are, their pre-existing attitudes towards AI. However, based on our studies, we can only speculate about this. Further research is needed to explore the mechanism underlying this differential effect, and may especially benefit from measures of attitudes towards and trust in AI-based writing assistants.

Moreover, user acceptance appears to be almost unaffected by the valence of the recommended emotional tone. Against our prediction, neither behavioral nor self-report acceptance was greater for recommendations of a positive emotional tone than for recommendations of a negative one. In fact, in Study 2, behavioral acceptance was significantly higher for negative recommendations, although this effect was singular. This contradicts previous research by Arnold et al. (2018), which found users preferring a positively biased predictive-text writing assistant, but may be explained by the employed paradigms: While Arnold et al. (2018) used a non-competitive task in their research, the present research used a competitive task. The current findings therefore suggest that users' preference for positively versus negatively biased systems may depend on the interaction context in which the system is used. Alternatively, the scenario paradigm used in the present research may not have been sufficiently realistic for impression management effects to appear, as participants were aware they were not actually interacting with another person. In this case, the competitive framing of the task would have favored the goal of achieving a good deal and the related communication of negative emotions. It seems unlikely that this effect would generalize to other, less competitive contexts, given that the present research specifically examined negotiations as a context in which, according to theory (Van Kleef et al., 2010), users showing a greater acceptance for negative recommendations could be justified. However, the fact that acceptance did not significantly differ between the recommended emotional tones in all but one case does not provide a strong basis for either conclusion.

Instead, the lack of significant effects could suggest that additional methodological considerations are necessary when examining this type of recommendation. For instance, the phrasing of the recommendations in the present research was rather subtle ('sound friendly' vs. 'sound assertive') compared to other possible phrasings (e.g., 'sound as happy as possible' vs. 'sound as angry as possible'). Participants' perceptions of what kind of change the recommendation suggested may have therefore not differed strongly enough between conditions to result in observable differences in acceptance. In the same vein, it cannot be ruled out that the difference between AI-based recommendations that included highlights and AI-based recommendations that did not (H1b) remained non-significant simply because of the way the highlights were implemented in the present research. The one-word highlights may have failed to increase the recommendations' functionality if participants disagreed about the highlighted section's relevance or considered the highlighted section too small to be applicable. In these cases, users may have even perceived revisions to require more effort if they needed to work around the highlights. This would have likely hindered acceptance, as previous research suggests a positive link between the (perceived) ease of using a system and acceptance (cf. Kelly et al., 2023). As the current research did not include measures for perceived ease of use (or related concepts), we can only speculate about this. Still, recommendations that provide detail with different forms of highlighting (e.g., entire sections) or even alternative phrasings may produce different effects.

Recipients' Reactions

As predicted, users' message revisions significantly increased recipients' perceptions of the recommended emotional tone, demonstrating that users generally followed the assistant's instructions. This finding gives increased relevance to the previously proposed positivity bias in commercial writing assistants (Arnold et al., 2018). Coupled with the fact that acceptance was not strongly dependent on the recommended emotional tone, it lends support to the notion that the use of emotionally biased writing assistants could lead to similarly biased messages in interpersonal online interactions. Importantly, the current findings also support the notion that such changes to the emotional tone of messages can have downstream effects on interpersonal interaction outcomes. In line with our prediction and the EASI-Model's assumptions (Van Kleef et al., 2010), messages revised after positive recommendations were not only perceived as more positive but also elicited more favorable interpersonal impression ratings than original messages. However, AI-functionalities did not moderate these effects. While more favorable interpersonal impressions may generally be considered a desirable interaction outcome, findings for recommendations of a negative emotional tone show that the downstream effect of writing-recommendations for emotional tone may not always be as desirable or as expected. Although messages revised after these recommendations were rated as more negative than original messages, they did not elicit a higher willingness to

make concessions from recipients. Instead, recipients reported a higher willingness to make concessions after messages they perceived as less negative. This finding was unexpected, as it seemingly contradicts the EASI-Model's assumptions (Van Kleef et al., 2010) and previous findings of negative communicated emotions (i.e., anger) eliciting higher concessions in negotiations (Steinel et al., 2008; Van Kleef et al., 2004).

However, it may simply demonstrate the difficulty of eliciting a desirable effect on concessions with recommendations of a negative emotional tone due to it being more prone to moderating influences than the effect of positive communicated emotions on interpersonal impression. According to the EASI-Model, the effect of negative communicated emotions on concessions in competitive negotiations is mediated by strategic inferences (e.g., inferring an unwillingness to accept a bad offer from expressions of anger) and may disappear or even reverse if recipients' reactions are guided by affective reactions rather than strategic inferences (Van Kleef et al., 2010). The likelihood of strategic inferences guiding recipients' reactions is proposed to decrease if the communicated emotion is perceived as inappropriate, for instance, due to misplaced intensity, perceived inauthenticity, or non-adherence to cultural or group norms (Van Kleef, 2014). These factors were not taken into consideration in the recommendations made in the current research, and given how highly context-dependent and thus not easily incorporated into automated writing recommendations these factors are, it is at least doubtful that real-world writing assistants would consider them. As a result, messages revised with these recommendations could more likely be perceived as more inappropriate by recipients, leading to fewer strategic inferences and negative downstream effects on concession behavior. Similarly, the EASI-model also proposes that recipients' reactions are less likely to be guided by strategic inferences if recipients are not sufficiently motivated to draw inferences, that is, if recipients' epistemic motivation is low. In the current research, this motivation may not have been sufficiently given, since recipients did not actually need to reach an agreement with their negotiation partner. The toughness communicated by a negative emotional tone may therefore not have posed a realistic threat to recipients' goals. As the current research did not include measures for perceived appropriateness or epistemic motivation, these considerations must remain speculative. Still, they illustrate that the current findings do not necessarily contradict the EASI-model's assumptions but instead highlight the need for future research addressing these potential moderators.

Finally, the fact that AI-based recommendations did not elicit stronger expressions of the recommended emotional tone than non-AI recommendations suggests that users' ability to revise messages in the desired direction successfully is not (chiefly) dependent on their ability to apply a given recommendation to their original message and identify relevant message parts. Although this finding was unexpected, it aligns with the fact that users' acceptance was also not consistently affected by the highlighting functionality. The factors that may have limited the functionality and therefore acceptance proposed above may also have limited their practical usefulness for message revisions, especially for recommendations of a negative emotional tone. For instance, given that the highlights in the present research addressed only a single word, the emotion communication in revised messages may have been rendered unconvincing for recipients if writers took a minimal approach by changing only what appeared to be strictly necessary, as indicated by the highlight.

Limitations, Strengths, and Future Research

Beyond the methodological considerations mentioned above, there are also some more general limitations to the current findings. The limited realism of the employed scenario paradigm may limit the external validity of findings. Similarly, it remains unclear whether the current findings would generalize to other negotiation contexts than the specific example (the terms of a bicycle repair) described in the employed scenario. Moreover, to ensure sufficient performance of our AI, as well as real-world applicability of the negotiation scenario's context for all participants, the samples for all three studies were convenience samples of native English speakers living in the UK. Given that the use of and responses to communicated emotions are culturally moderated (Safdar et al., 2009), future studies should replicate the present findings with more diverse samples. On the technical level, the emotion classifier may, in some cases, have produced misclassifications leading to recommendations of the same rather than the opposite emotional tone in the AI-based conditions of Study 1 and 2. However, the number of these cases should be negligible, given the classifier's accuracy scores. Moreover, since the current research was conducted, the field of AI-agents has shifted to decoder-only Transformer models that focus on text generation. With such text generation methods showing impressive zero-shot predictive capabilities, a straightforward methodological choice would now be to replace the classifier and the explainer module with such autoregressive methods, configured via prompting. However, this approach would not change the research question or design of the

present paper, and we are confident that the findings are robust regarding changes to the backend. In fact, the present research's implementation has the advantage of being reproducible and understandable without substantial effort, and of allowing the detection of potential model biases through its explainer, which the use of a large language model behind an API might not offer.

On the other hand, the current research also has key strengths. It examines the impact of a type of recommendation already used in real-life online interactions but has rarely been addressed in psychological research. It thus addresses a clear gap in research, even though the lack of previous findings may make it more challenging to contextualize unexpected or mixed findings in the present studies. It addresses this gap with a structured approach. Taken together, the three presented studies examined the effect of AI-based recommendations for messages' emotional tone on the interaction between message senders and message recipients, with a focus on specific AI-based functionalities. This approach has several strengths. Focusing on both message senders and recipients in our experiments allows for a more complete insight into the recommendations' impact on online interactions, while the focus on specific recommendation characteristics allows for a more detailed discussion of the potential drivers of this impact. The experimental data collected using a self-developed AI provides strong support for causal conclusions in a field of research that often relies on existing commercial writing assistants with an unknown number of potential confounding variables. Especially the finding of AI-based functionalities increasing users' acceptance of recommendations for messages' emotional tone can reasonably be expected to generalize beyond the context of negotiations, as it was consistently unaffected by the valence of the recommended emotional tone. Additionally, all studies were preregistered.

Still, many questions regarding the impact of AI-based recommendations for emotional tone remain unexplored, offering ample avenues for future research. First, future research could use more realistic paradigms, including actual or simulated negotiations, and address different negotiation objects to ensure the generalizability of the current findings. Future studies could also examine the potential impact of various moderating variables for both user acceptance (e.g., perceived ease of use, trust, or attitudes towards AI) and recipients' reactions (e.g., epistemic motivation or factors influencing the perceived appropriateness of communicated emotions). They could additionally consider the potential effect of external variables like message senders' and recipients' incidental moods, prior experience with negotiations, or cultural negotiation styles. Alternatively, future studies could focus on recommendations that promote other emotional tones, or could directly compare the effects of recommendations that promote a positive versus negative emotional tone in competitive and cooperative contexts, and verify whether user acceptance is context-dependent. Within all these examinations, pinpointing minor, yet potentially important (and confounding) recommendation characteristics (e.g., how highlights are implemented or the exact phrasing of a recommendation) a priori may pose a challenge, given the scarcity of previous research. Studies examining the impact of different feature implementations would therefore also be particularly valuable.

Theoretical and Practical Implications

The current research demonstrated that recommendations regarding communication can have a severe impact on communication and its outcomes. Users revised messages in line with (AI-based) recommendations, and revised messages affected recipients' perception of the messages' emotional tone and the sender, as well as their willingness to make concessions. In other words, there is strong potential that AI-based writing assistants assert a substantial influence on the course of an interaction. Depending on the pre-settings of the respective system or the user's personal settings, which can take various forms, a social interaction can take very different routes. Thereby, a new factor has entered the stage of social interaction, requiring a change in theorizing about mediated communication.

From a practical perspective, the current findings imply that writing assistants are a resource in strategic communication. They can help make desired impressions and potentially also reach certain interaction goals. Within certain communication, this can be advantageous. In the long run, the effect of the use of writing assistants will most likely depend on a host of factors, such as the authenticity of the tone and the consistency across interactions. In other words, what has a beneficial potential will not necessarily result in long-term positive effects. Research on the downstream consequences of writing assistant use would be desirable to understand its applied and theoretical consequences.

Conclusion

Using an experimental approach with a self-developed AI, the current research examined the effect of AI-based writing assistants for messages' emotional tone on interpersonal interactions in written online negotiations, with a focus on message senders' acceptance of the assistant and its recommendations and recipients' reactions to the produced messages. Results indicate that AI-functionalities do increase users' acceptance, especially at the behavioral level, whereas the valence of the recommended emotional tone is less important for acceptance. On the side of recipients, results indicate that message revisions generally do increase perceptions of the respective recommended emotional tone and that this can lead to more favorable interpersonal impressions if a positive emotional tone is recommended. Taken together, these results show that AI-features in writing assistants can increase user acceptance and that message revisions generally tend to change the emotional tone of interpersonal messages in the recommended direction, which, in turn, affects recipients' impressions of the message sender. However, the implementation of revisions once recommendations are accepted, and the resulting messages' effects on recipients do not seem to depend on whether recommendations are AI-based.

Footnotes

[1] Hypotheses 1a and 1b were preregistered as two separate hypotheses but combined in the manuscript for readability. Two additional preregistered hypotheses for Study 2 are reported in the Appendix for readability.

[2] The functionality factor was preregistered with two separate (quasi-experimental post-hoc) non-AI conditions meant to control for the fit between the recommended emotional tone and the emotional tone of participants' original message. Upon review, this produced an ill-fitted experimental design. The two conditions were thus combined in the main manuscript. For details and analyses including the two separate conditions please see the Appendix.

[3] The two separate neutral categories arose from a pre-processing mistake in the AI's training data (please refer to the AI-description in the Appendix for details). Excluding participants whose message had received one of these classifications from analyses in Study 1 and 2, did not reveal a different pattern of effects.

[4] As in Study 1, functionality was preregistered to include two separate non-AI conditions (i.e., four groups, see footnote 2). The sample size calculation was based on the preregistration.

[5] CPU 2 x Intel(R) Xeon(R) Gold 5218B, RAM 256GB, GPU Nvidia Tesla V100.

Conflict of Interest

The authors have no conflicts of interest to declare.

Use of AI Services

The authors declare that they have not used any AI services to generate or edit any part of the manuscript, data, or stimuli, beyond the self-developed AI integral for the experimental manipulations as stated in the manuscript.

Data Availability Statement

All presented studies were preregistered in full following the standard questions on aspredicted. The authors agree to share their materials, data, and analytic methods with other researchers. The materials are provided in the Appendix. Preregistrations, data, codebooks, and analysis syntaxes can be retrieved from https://researchbox.org/2774.

Authors' Contribution

Josephine Hagedorn: conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, writing—original draft, writing—review & editing. **Roman Klinger:** software, methodology, writing—review & editing. **Kai Sassenberg:** conceptualization, formal analysis, funding acquisition, methodology, project administration, supervision, writing—review & editing.

Acknowledgment

This article is based on research conducted as part of the first author's PhD thesis: 'Hagedorn, J. C. (2025). *The Impact of AI-based Writing Recommendations for Emotional Tone on Interpersonal Online Interactions* [Doctoral dissertation, University of Tübingen]. TOBIAS-lib.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Apple Inc. (n.d.). *Apple Intelligence*. Apple. https://www.apple.com/apple-intelligence/

Arnold, K. C., Chauncey, K., & Gajos, K. Z. (2018). Sentiment bias in predictive text recommendations results in biased writing. In *GI '18: Proceedings of the 44th Graphics Interface Conference* (pp. 42–49). Association for Computing Machinery. https://doi.org/10.20380/GI2018.07

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Davis, F. D. (1993). User acceptance of information technology: System characteristics, user perceptions and behavioral impacts. *International Journal of Man-Machine Studies*, *38*(3), 475–487. https://doi.org/10.1006/imms.1993.1022

Gieselmann, M., Hagedorn, J., & Sassenberg, K. (2024). Do perceived benefits compensate for low provider trustworthiness in disclosure decisions? An experimental investigation. *Journal of Media Psychology: Theories, Methods, and Applications, 37*(6), 376–387. https://doi.org/10.1027/1864-1105/a000440

Grammarly Inc. (n.d.). *Our Features | Grammarly*. Grammarly. https://www.grammarly.com/features

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society*, *22*(1), 70–86. https://doi.org/10.1177/1461444819858691

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89–100. https://doi.org/10.1093/jcmc/zmz022

Harker, L., & Keltner, D. (2001). Expressions of positive emotion in women's college yearbook pictures and their relationship to personality and life outcomes across adulthood. *Journal of Personality and Social Psychology*, *80*(1), 112–124. https://doi.org/10.1037/0022-3514.80.1.112

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, *40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Help Scout. (n.d.). *AI-powered customer service by Help Scout*. Help Scout. https://www.helpscout.com/ai-features/

Hillebrandt, A., & Barclay, L. J. (2017). Comparing integral and incidental emotions: Testing insights from emotions as social information theory and attribution theory. *Journal of Applied Psychology*, *102*(5), 732–752. https://doi.org/10.1037/apl0000174

Hohenstein, J., & Jung, M. (2018). AI-supported messaging: An investigation of human-human text conversation with AI support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)* (Article LBW089). Association for Computing Machinery. https://doi.org/10.1145/3170427.3188487

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., & Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, *13*(1), Article 5487. https://doi.org/10.1038/s41598-023-30938-9

Homan, A. C., Van Kleef, G. A., & Sanchez-Burks, J. (2016). Team members' emotional displays as indicators of team functioning. *Cognition and Emotion*, *30*(1), 134–149. https://doi.org/10.1080/02699931.2015.1039494

Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*, Article 101925. https://doi.org/10.1016/j.tele.2022.101925

Kopalasingam, M, & Greene, J. (2024, January 10). *Introducing AI assist for better, faster responses—Help Scout.* Inside Help Scout. https://www.helpscout.com/blog/introducing-ai-assist/

Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, *7*(4), 730–735. https://doi.org/10.1037/1528-3542.7.4.730

Levine, E. E., Barasch, A., Rand, D., Berman, J. Z., & Small, D. A. (2018). Signaling emotion and reason in cooperation. *Journal of Experimental Psychology: General*, *147*(5), 702–719. https://doi.org/10.1037/xge0000399

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv. https://doi.org/10.48550/ARXIV.1907.11692

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, Article 121390. https://doi.org/10.1016/j.techfore.2021.121390

Mieczkowski, H., Hancock, J. T., Naaman, M., Jung, M., & Hohenstein, J. (2021). AI-mediated communication: Language use and interpersonal effects in a referential communication task. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), Article 17. https://doi.org/10.1145/3449091

OpenAI. (n.d.). *ChatGPT*. OpenAI. https://openai.com/de-DE/chatgpt/overview/

Pitardi, V., & Marriott, H. R. (2021). Alexa, she's not human but... Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing*, *38*(4), 626–642. https://doi.org/10.1002/mar.21457

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Safdar, S., Friedlmeier, W., Matsumoto, D., Yoo, S. H., Kwantes, C. T., Kakai, H., & Shigemasu, E. (2009). Variations of emotional display rules within and across cultures: A comparison between Canada, USA, and Japan. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *41*(1), 1–10. https://doi.org/10.1037/a0014387

Sinaceur, M., & Tiedens, L. Z. (2006). Get mad and get more than even: When and why anger expression is eVective in negotiations. *Journal of Experimental Social Psychology*, *42*(3), 314–322. https://doi.org/10.1016/j.jesp.2005.05.002

Steinel, W., Van Kleef, G. A., & Harinck, F. (2008). Are you talking to me?! Separating the people from the problem when expressing emotions in negotiation. *Journal of Experimental Social Psychology*, *44*(2), 362–369. https://doi.org/10.1016/j.jesp.2006.12.002

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Van Kleef, G. A. (2014). Understanding the positive and negative effects of emotional expressions in organizations: EASI does it. *Human Relations*, *67*(9), 1145–1164. https://doi.org/10.1177/0018726713510329

Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2004). The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology*, *86*(1), 57–76. https://doi.org/10.1037/0022-3514.86.1.57

Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2010). An interpersonal approach to emotion in social decision making: The emotions as social information model. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 42, pp. 45–96). Elsevier. https://doi.org/10.1016/S0065-2601(10)42002-X

Walther, J. B., & D'Addario, K. P. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review*, *19*(3), 324–347. https://doi.org/10.1177/089443930101900307

Wigboldus, D. H. J., Semin, G. R., & Spears, R. (2000). How do we communicate stereotypes? Linguistic bases and inferential consequences. *Journal of Personality and Social Psychology*, *78*(1), 5–18. https://doi.org/10.1037/0022-3514.78.1.5

Ye, Z., Lelieveld, G.-J., & Van Dijk, E. (2025). Evaluating negotiators who deceptively communicate anger or happiness: On the importance of morality, sociability, and competence. *Journal of Business Ethics*, *199*(4), 799–817. https://doi.org/10.1007/s10551-024-05824-7

Appendix

Materials

The following are the main materials used in the studies. Across these materials (except in the lists of the measures), square brackets indicate additional notes from the authors and '--' indicates a new page.

Study 1

In the following, you will read a scenario describing the course of an online negotiation about a bicycle repair. We will ask you to imagine yourself in this situation as well as you can and formulate a message to your described negotiation partner. Next, your message will be checked by an AI-based negotiation assistant, which will make a recommendation on how to potentially reach a better negotiation outcome. You'll then have the option to edit your message.

--

Please read the following scenario carefully and try to imagine yourself in the situation to the best of your ability:

You have been after the bicycle of your dreams for a while. Now, you were finally able to buy it. It is used, and sadly it is in no good condition. However, according to a friend who knows a lot about bikes, the repairs should be manageable if you are a tough negotiator. The price of the repair is essential to you since you have a restricted budget. However, it is also important to you that there is a warranty for the repair with a longer duration than the legally ensured six months. You would also like the repair shop to make an effort to get the needed parts for the repair quickly since you are obviously looking forward to riding your new bicycle. With this in mind, you contact the only repair shop in your area that has the appropriate know-how for the repair.

--

Against this background, the situation develops as follows:

After some research, you're reasonably sure that the repairs should cost no more than 160 £ with a delivery time of 5 days and a 24-month warranty. Trying to be a tough negotiator, you send an email to the shop requesting a repair that has a slightly longer delivery time for the parts (3 days more), but is 30 £ cheaper, and has 3 months more warranty compared to the results of your research. In their reply, the shop offers a repair which is 120 £ more expensive, has 13 days longer delivery time, and 15 months less warranty than your request. After some back-and-forth, the shop's offer has considerably improved (compared to your initial request, it is now 30 £ more expensive, with the same delivery time for the parts and 3 months less warranty). You get the feeling that this is close to the optimum you can achieve and that, depending on your next steps, you could get a rather good deal out of them.

Now, you may write a message to the repair shop, telling them how you feel about their offers and behaviour up this point. Your message should be exactly one sentence.

What message do you want to send to the repair shop to let them know how you feel about their offers up to this point?

Please try to formulate only one full sentence but answer spontaneously. There are no right or wrong answers.

One moment please.

[Depending on the experimental condition one of the following recommendations was presented]

[AI-based non-detailed or non-AI recommendation of a negative emotional tone:]

The Assistant's Recommendation

'[Participant's original message]'

To increase your chances of getting a good offer from your negotiation partner, your message should sound assertive.

[AI-based non-detailed or non-AI recommendation of a positive emotional tone:]

The Assistant's Recommendation

'[Participant's original message]'

To increase your chances of making a good impression on your negotiation partner, your message should sound friendly.

The Assistant's Recommendation

'[Participant's original message]'

To increase your chances of making a good impression on your negotiation partner, your message should sound friendly.

[Detailed AI-based recommendation of a negative message emotionality:]

The Assistant's Recommendation

'[Participant's original message]'

To increase your chances of getting a good offer from your negotiation partner, your message should sound assertive.
The highlighted word makes your message sound less assertive.

[Detailed AI-based recommendation of a positive message emotionality:]

The Assistant's Recommendation

'[Participant's original message]'

To increase your chances of making a good impression on your Negotiation partner, your message should sound friendly. The highlighted word makes your message sound less friendly.

--

Do you want to change your message? [Behavioral Acceptance]

o Yes o No

--

[If 'Yes' was selected:]

Your original message was:

'[Participant's original message]'

Please type in your revised message:

--

Thank you. The scenario-based part of the study is now completed. In the next step, we would like to ask you a few questions about your experience during the task.

Measures in Study 1

The following are all measures included in the study, apart from items regarding basic information (e.g., demographics).

Main measures:

- Behavioral acceptance (1 item, 'Do you want to change your message', dichotomous 'yes', 'no')
- Self-report acceptance (2 items, 'I would use the negotiation assistant in the future' [use intentions], 'I think the negotiation assistant is useful' [perceived usefulness], and use intentions, both 5-point Likert-scale, 1 'Disagree Completely', 5 'Agree Completely').

Exploratory Measures:

- Perceived sensibility of the assistant's recommendation (1 item, 'The negotiation assistant's recommendation was sensible.', 5-point Likert-scale, 1 'Disagree Completely', 5 'Agree Completely')
- Perceived sensibility of word marked by assistant (1 item, only in AI-based functionality condition, 'The negotiation assistant's recommendation on which word to change was sensible.', 5-point Likert-scale, 1 'Disagree Completely', 5 'Agree Completely')
- Participants' goals while writing their original message (1 item, 11-point bipolar scale, 'When I wrote my initial message, I tried to…', 0 'Make a good impression', 10 'Get the best deal')
- Changes in emotional tone between participants original and revised messages (as classified by the AI used for the writing assistant in the study; post hoc)

Attention Checks:

- Attention Check 1: 'This is an attention check, please select '5 Agree completely', passed if instruction was followed

- Attention Check 2: 'I swim across the Atlantic ocean to work every day.', passed if 1 'Completely disagree' or 2 'Disagree' was selected.

Study 2

The materials for Study 2 were the same as the materials for Study 1, with the only exception being that one exploratory item regarding participants' goals while writing their message was presented directly after participants wrote their message instead of after the scenario task. Thus instead of the wait screen in Study 1, Study 2 displayed the following before the assistant's recommendation was displayed:

While your message is being processed, we would like to ask you about the message you just wrote.

Please adjust the following slider to a position you feel most accurately presents your goals while writing your message.

When I wrote my message, I tried to…

Make a good impression 0 1 2 3 4 5 6 7 8 9 10 Get the best deal

Measures in Study 2

All measures in Study 2 were the same as in Study 1, except for the item regarding the perceived sensibility of the word marked by the assistant, which was omitted in Study 2.

Study 3

For the full list of presented messages, please refer to the data (some examples are provided in Table A1 below).

Table A1. *Examples for Message Presented in Study 3.*

Message Origin Condition		Message Version	
Functionality	Emotional Tone	Original	Revised
Non-AI	Positive	*Thanks for all your help with this so far but can you do any better on the costs and time?*	*Hi there :) thanks for taking the time with this and all your help so far - can you do any better on the cost and time?*
Non-AI	Positive	*Would it be possible for the price to be reduced by £15 due to the 3 months less warranty?*	*Thank you so much for your help so far, I very much appreciate the offer but would it be possible to reduce the price by £15 due to the 3 months less warranty?*
NonAI	Negative	*Im a bit happy but would like the warranty to be the same.*	*Either you reduce the price of parts or increase the warranty or i will go elsewhere.*
AI	Positive	*Frankly, your offers so far have been almost unacceptable and I still feel like I am being overcharged for repairs.*	*Frankly, your offers so far have not been meeting my expectations and I still feel like I am being overcharged for repairs but I appreciate your effort to reach a satisfactory middle ground*
AI	Negative	*I would be more than happy to accept this offer if we can stick to the original £120 providing the delivery time is longer.*	*I will accept your offer of £120 providing the delivery time is now longer.*

In the following, you will read a scenario describing the course of an online negotiation about a bicycle repair. We will ask you to imagine yourself in this situation as well as you can and rate several different possible messages from your negotiation partner. The scenario will be the same across messages but will remain visible to you.

--

Please imagine you work in a bicycle shop and offer bicycle repairs. You have standard offers for these repairs but do have some wiggle room regarding the prices and warranty periods you offer. You could also make an effort to accelerate the delivery time. One day, you receive an email from a customer, requesting a repair for 130 £, with

a 27-month warranty period and delivery time of 8 days. As that is a little steep, you answer with your standard offer for the kind of requested repair: 250 £, 12-month warranty period, and a delivery time of 21 days. However, the customer is keen to negotiate and after some back-and-forth, you've made some concessions. You last offered the repair for 160 £, with a warranty period of 24 months, and an 8-day delivery time.

Next, you receive the following message from the customer:

The customer's message:

'*[Message from study 1 or 2; repeated 10 times]*'

Please indicate how much you agree with the following statements.

Measures in Study 3

Next to attention checks (attention check 1 and 2 were passed if the instruction was followed and attention check 3 was the same as attention check 2 in Study 1 and 2) and items regarding basic information (e.g. demographics), the following was assessed:

- Willingness to make Concessions (1 item, 'I would make concessions after receiving this message')
- Positive Impression of the message's sender (1 item "I would enjoy working with this customer in the future')
- Perceived positive emotional tone (3 items, 'This message expresses [satisfaction/happiness/friendliness]')
- Perceived negative emotional tone (3 items, 'This message expresses [assertiveness, irritation / anger]')

All 9-point Likert-scales, 1 'Disagree Completely – 9 'Agree Completely'.

Deviations from Preregistrations

All Studies

Table A2. *Changes in Wording of Hypotheses.*

	Wording in Pre-registration		Wording in Manuscript
H1	The acceptance of participants who receive an AI-based recommendation will be higher than the acceptance of participants who receive a recommendation that is not AI-based.	H1	Participants will show more acceptance (a) when recommendations are AI-based compared to not AI-based and (b) when AI-based recommendations highlight relevant message parts for revisions compared to when they do not.
H2	When participants only receive the AI-based adaptive instruction regarding the emotional tone to be achieved, acceptance will be lower than when the AI additionally marks the most important word for the undesired emotion expressed in the original message.		
H3	The acceptance of participants who receive a recommendation promoting a more positive message emotionality will be higher than the acceptance of participants who received a recommendation promoting a more negative message emotionality.	H2	Participants will be more likely to accept a recommendation promoting a more positive emotional tone for a message than a recommendation promoting a more negative emotional tone for a message.
H1	Original messages will be rated less in line with the emotionality and intention implied by the recommendation than revised messages – more so in the case of messages written with a AI-based recommendation than in case of messages written with a non-AI-based recommendation.	H3	Recipients will rate revised messages more in line with the emotional tone and communication goal implied by the recommendation the messages were revised with than original messages – more so for messages revised with an AI-based recommendation than for those revised with a non-AI-based recommendation.

To improve readability, the term 'recommendation target emotionality' describing one of the experimental factors in the pre-registration, was changed to 'emotional tone' for Study 1 and 2 and to 'recommended emotional tone'

for Study 3, and the term 'recommendation type' describing another experimental factor was changed to 'functionality' in Study 1 and 2 and to 'recommendation functionality' in Study 3.

Hypotheses 1 and 2 in the preregistrations for study 1 and 2 were combined into one in the manuscript and Hypothesis 3 in the preregistrations was consequently renamed into Hypothesis 2 in the manuscript. Hypothesis 3 in the manuscript was labelled as H1 in the pre-registration for study 3. Additionally, the wording of the hypotheses in the pre-registrations was simplified in the manuscript for better readability, while taking care that no changes to meaning occurred (see Table A2 for all changes in wording).

Study 1

Study 1 was preregistered as two separate studies (separated between AI-based and non-AI recommendations) that may, however, be considered one experiment. Since the data collection was only split into separate studies due to technical reasons, as pointed in the preregistration, they are simply reported as one study with separate data collections in the main text.

Due to a labelling error in the preprocessing of the AI's training data, the two neutral emotionality classifications were initially labelled as 'neutral' (instead of 'neutral$_A$') and 'sad' (instead of 'neutral$_B$'). The description of the recommended emotionality conditions in the preregistration still includes the old labels, whereas they have been updated in the manuscript. This was also the case for Study 2.

Additionally, the scale for an exploratory item regarding participants' goals while writing their message (not reported in the main manuscript) was preregistered to be anchored by 1 'Make a good impression', 10 'Get the best deal' but was instead anchored by 0 'Make a good impression', 10 'Get the best deal'. This change should have no consequences but is reported for completeness. This was also the case for Study 2.

We preregistered a separate ANOVA for self-report acceptance by recommendation type (2 groups: AI-detailed vs. non-detailed AI) and emotional tone (2 groups: positive vs. negative) for the comparison of detailed and non-detailed AI-based recommendations and AI non-detailed recommendations (H1 in preregistration, H1b in manuscript), however in the manuscript we test H1b with the same contrast analysis that is used to test H1a to improve clarity and readability. Results of this separate ANOVA (that only used the data from the AI-based functionality conditions, $n = 381$) are reported here. The critical main effect of functionality ('recommendation type' in preregistration) was not significant, $F(1,377) = 0.09$, $p = .762$, *part.* $\eta^2 = .000$, nor was there a significant main effect of the recommended emotional tone, $F(1,377) = 1.40$, $p = .238$, *part.* $\eta^2 = .004$. There was a significant interaction between the two factors, $F(1,377) = 4.00$, $p = .046$, *part.* $\eta^2 = .011$, as negative recommendations lead to higher self-report acceptance than positive recommendations only if participants were presented with the additional highlights (AI-based detailed recommendation), $p = .027$. There were no other significant differences, all $p > .1$.

Finally, as pointed out in the main manuscript, the main manuscript includes only one non-AI Condition in the functionality factor in Studies 1 and 2, while the preregistration included two separate non-AI conditions, resulting in four groups in the functionality factor (Variable AI_COND2 in the analyses syntax). These two conditions were quasi-experimental post-hoc conditions. They indicated whether the recommended emotional tone (randomly assigned for non-AI recommendations) had been the opposite to the one present in participants' original message (non-AI fitting condition), as was also the case for the AI-based recommendations, or whether the recommended emotional tone was the same as the one present in participants' original message (non-AI non-fitting). This was originally meant to control for the fit between the recommended emotional tone and participants' message (recommending a change in emotional tone vs. promoting the already present emotional tone), however, upon review, this produced an ill-fitted experimental design, as the 'AI-based Non-Detailed' condition and the 'non-AI fitting condition' presented the same to participants and were never to be directly compared. To address this, we combined the two non-AI conditions in the main manuscript but report the analyses testing H1a and H2 (H2 and H3 in the preregistration) with the preregistered conditions in the following. Descriptive statistics are reported in Table A3.

A two-way ANOVA for *self-report acceptance* by functionality (AI-based detailed vs. AI-based non detailed vs. non-AI fitting vs. non-AI non-fitting) and emotional tone (positive vs. negative) showed a significant main effect of functionality, $F(3,755) = 3.61$, $p = .013$, *part.* $\eta^2 = .014$. A follow-up contrast analysis with the critical contrast comparing both AI-conditions with both non-AI conditions (1 1 −1 −1) showed significantly higher self-report acceptance for AI-based recommendations than non-AI recommendations, $t(759) = 3.41$, $p < .001$, $d = .494$. The

orthogonal contrast for AI-detailed vs. AI non-detailed recommendations (1 -1 0 0) was non-significant, $t(759) = 0.08$, $p = .939$, $d = .008$. The second orthogonal contrast for non-AI fitting vs. non-AI non-fitting recommendations (0 0 1 −1) also remained non-significant, $t(759) = -0.19$, $p = .852$, $d = .02$. The main effect of emotional tone remained non-significant, $F(1,755) = 3.19$, $p = .075$, *part.* $\eta^2 = .004$. The interaction between the two experimental factors also remained non-significant, $F(3,755) = 1.54$, $p = .203$, *part.* $\eta^2 = .006$.

For *behavioral acceptance*, a logistic regression with emotional tone (positive vs. negative) and three contrasts for functionality (both AI-based conditions vs. both non-AI based conditions, AI-detailed vs. AI non-detailed condition, non-AI fitting vs. non-AI non-fitting condition: 0.5 0.5 −0.5 −0.5, 0.5 −0.5 0 0, 0 0 0.5 −0.5), as well as the interaction terms between the contrasts and emotional tone as predictors. Participants opted to change their message significantly more often after AI-based than after non-AI based recommendations, $B = .88$, $SE = .15$, $\beta = 2.40$, $p < .001$, regardless of which emotional tone was recommended, $B = .24$, $SE = .15$, $\beta = 1.27$, $p = .113$. The contrast between the AI-detailed and the AI non-detailed condition showed that participants' decision to change their message was significantly predicted by both (a) the detailedness of AI-based recommendation, $B = .46$, $SE = .21$, $\beta = 1.59$, $p = .030$, and (b) the interaction between the detailedness of AI-based recommendation and emotional tone, $B = -.46$, $SE = .21$, $\beta = .63$, $p = .032$. Participants opted to change their message more often in the AI-detailed condition than in the AI non-detailed condition and this was specifically the case when a negative emotional tone was recommended. Emotional tone did not predict behavioral acceptance, $B = -.07$, $SE = .08$, $\beta = 0.93$, $p = .360$. The interaction of emotional tone with the second orthogonal contrast, non-AI fitting vs. non-AI non-fitting, also remained non-significant, $B = .06$, $SE = .22$, $\beta = 1.07$, $p = .766$. The difference between the non-AI fitting and the non-AI non-fitting condition also remained non-significant, $B = -.01$, $SE = .22$, $\beta = 0.99$, $p = .960$.

Table A3. *Study 1: Descriptive Statistics for Acceptance Measures.*

Factor			Acceptance Measure	
Functionality	Emotional Tone	*n*	Self-report	Behavioral
			M (*SD*)	% Opted to Change Message
AI-detailed	Negative	112	3.58 (1.04)	67.0
	Positive	75	3.23 (1.08)	58.7
AI non-detailed	Negative	112	3.40 (1.11)	44.6
	Positive	82	3.49 (1.01)	58.5
Non-AI fitting	Negative	109	3.27 (1.03)	39.4
	Positive	90	3.04 (1.15)	32.2
Non-AI non-fitting	Negative	80	3.22 (1.00)	41.3
	Positive	103	3.16 (1.01)	31.1
Overall	Negative	413	3.38 (1.06)	48.7
	Positive	350	3.22 (1.07)	43.7
	Overall	763	3.31 (1.06)	46.4

Study 2

To facilitate a more consistent reporting of results across Studies 1 and 2, we report a follow-up contrast analysis testing Hypotheses 1a and 1b for self-report acceptance in Study 2, although this analysis was only preregistered in case of a significant main effect of recommendation functionality, which was not given in Study 2.

As in Study 1, we preregistered a separate ANOVA for self-report acceptance by recommendation type (2 groups: AI-detailed vs. non-detailed AI) and emotional tone (2 groups: positive vs. negative) to test H1b (H1 in preregistration), but in the manuscript H1b is tested with the same contrast analysis that is used to test H1a to improve readability. Results of the preregistered ANOVA (that only used the data from the AI-based functionality conditions, $n = 379$) are reported here. The critical main effect of recommendation type was not significant, $F(1,375) = 0.02$, $p = .890$, *part.* $\eta^2 = .000$. The main effect of recommended emotionality and the interaction between the two factors also remained non-significant, $F(1,375) = 0.77$, $p = .379$, *part.* $\eta^2 = .002$, and $F(1,375) = 0.50$, $p = .482$, *part.* $\eta^2 = .001$, respectively.

As was also the case for Study 1, Study 2 has only one non-AI Condition in the functionality factor in the main manuscript, whereas the preregistration included two separate non-AI conditions, that is four groups in the

functionality factor (Variable AI_COND2 in the data). These two separate non-AI conditions were the same as in Study 1 and were combined in the main manuscript for the same reasons. The analyses testing H1a and H2 (H2 and H3 in the preregistration) with four groups in the functionality factor are reported in the following. The descriptive statistics for the analyses are reported in Table A4. We preregistered a two-way ANOVA for self-report acceptance by functionality (4 groups: detailed AI vs. non-detailed AI vs. non-AI fitting vs. non-AI non-fitting) and emotional tone (positive vs. negative). The main effect of functionality remained non-significant, $F(3,761) = 1.41$, $p = .240$, *part.* $\eta^2 = .006$. A follow-up contrast analysis with the critical contrast comparing both AI-conditions with both non-AI conditions (1 1 −1 −1) showed no significant difference in self-report acceptance between AI-based and non-AI recommendations, $t(765) = 1.83$, $p = .068$, $d = .264$. The orthogonal contrast for AI-detailed vs. AI non-detailed recommendations (1 −1 0 0) also remained non-significant, $t(765) = 0.28$, $p = .769$, $d = .031$. The second orthogonal contrast comparing non-AI non-fitting recommendations with non-AI fitting recommendations (0 0 1 −1) remained non-significant, $t(765) = -0.73$, $p = .468$, $d = .071$. The main effect of emotional tone was significant, $F(1,761) = 5.34$, $p = .021$, *part.* $\eta^2 = .007$. The self-report acceptance of participants in the negative condition was significantly higher than that of participants in the positive condition. There was no significant interaction between the two experimental factors, $F(3,761) = 0.66$, $p = .577$, *part.* $\eta^2 = .003$.

Effects for behavioral acceptance, were examined with the same logistic regression as in Study 4.1 with emotional tone (positive vs. negative) and three contrasts for functionality (both AI-based conditions vs. both non-AI based conditions, AI-detailed vs. AI non-detailed condition, non-AI fitting vs. non-AI non-fitting condition: 0.5 0.5 −0.5 −0.5, 0.5 −0.5 0 0, 0 0 0.5 −0.5), as well as the interaction terms between the contrasts and emotional tone as predictors. Participants opted to change their message significantly more often after both AI-based recommendations than both non-AI based recommendations, $B = .70$, $SE = .15$, $\beta = 2.00$, $p < .001$. This was independent of the recommended emotional tone as there was no significant interaction, $B = -.17$, $SE = .15$, $\beta = .84$, $p = .259$. The effect of the first orthogonal contrast comparing the AI-detailed and AI non-detailed conditions remained non-significant, $B = .06$, $SE = .21$, $\beta = 1.07$, $p = .764$, independent of which emotional tone had been recommended, $B = -.19$, $SE = .21$, $\beta = .83$, $p = .368$. The effect of the second orthogonal contrast comparing the non-AI fitting and the non-AI non-fitting condition was significant, $B = .60$, $SE = .21$, $\beta = 1.82$, $p = .005$. Participants opted to change their message more often in the non-AI fitting condition, independent of which emotional tone was recommended, $B = .124$, $SE = .212$, $\beta = 1.132$, $p = .559$. Results did not show a significant effect of the recommended emotional tone, $B = -.120$, $SE = .075$, $\beta = 0.887$, $p = .108$.

Table A4. *Study 2: Descriptive Statistics for Acceptance Measures.*

Factor			Acceptance Measure	
Functionality	Emotional Tone	*n*	Self-report	Behavioral
			M (SD)	% Opted to Change Message
AI-detailed	Negative	112	3.53 (0.95)	65.2
	Positive	73	3.36 (1.09)	50.7
AI non-detailed	Negative	103	3.44 (1.06)	59.2
	Positive	91	3.42 (1.00)	53.8
Non-AI fitting	Negative	111	3.40 (1.01)	46.8
	Positive	81	3.26 (1.14)	48.1
Non-AI non-fitting	Negative	82	3.47 (1.09)	35.4
	Positive	116	3.26 (1.11)	31.0
Overall	Negative	408	3.46 (1.02)	52.7
	Positive	361	3.28 (1.09)	44.6
	Overall	769	3.38 (1.06)	48.9

Moreover, in addition to the ones reported in the manuscript, two more detail-oriented hypotheses were pre-registered for study 2, based on the finding of study 1. Results are reported here for improved consistency in hypotheses across studies.

*H4***:** *In the detailed AI-condition (but not in the non-detailed AI condition) self-report acceptance will be higher when a negative emotionality is recommended than when a positive emotionality is recommended.* This hypothesis was pre-registered to be tested with simple effect analyses resolving the interaction effect of recommendation type and

recommended emotionality in the ANOVA testing H1. However, results showed no significant interaction of the two experimental factors, p = .482. For descriptive values, see Table 2 in the manuscript.

H5: When a negative message emotionality is recommended (but not when a positive emotionality is recommended), behavioural acceptance in the detailed AI-condition will be higher than in the non-detailed AI condition.

This hypothesis was tested using only data from the AI-based recommendation type conditions in a X^2. Test for behavioral acceptance by recommendation type (2 groups: detailed AI vs. non-detailed AI) and recommendation target emotionality (2 groups: positive vs. negative). Results showed no significant differences between the two recommendation types if a negative emotionality was recommended, p = .368, nor if a positive one was recommended, p = .162. For descriptive values, see Table 2 in the manuscript.

Study 3

For clarity, the term 'message type' describing one of the experimental conditions in the preregistration was changed to 'message version'.

Additionally, we pre-registered two separate contrast analyses for each recommended emotionality condition with the same dependent variables as in the main analysis (positive [negative] emotionality ratings and interpersonal impression [willingness to make concessions] ratings) to test for baseline differences. The focal contrast (unrevised 2, original AI −1, original non-AI −1) served to test whether unrevised messages differ from messages that were later revised. The second contrast (0 1 −1) serves to test the difference between original messages (later revised) in the AI and the non-AI recommendation type condition. Both contrasts were entered as within factors into the analyses (without an interaction between the two; for descriptive values see Table A5). Revised messages were not included in this analysis.

For the negative recommended emotionality condition, neither the first nor the second contrast was significant for negative emotionality ratings, p = .874 and p = .915 respectively, or ratings of concession intentions, p = .924 and p = .787, respectively. In the positive recommended emotionality condition, the focal contrast between original messages that were later revised and unrevised messages was significant for positive emotionality ratings, p = .028, and interpersonal impression ratings, p = .008. For both dependent variables, unrevised messages were rated higher than messages that were later revised. Meanwhile, the second contrast between original messages in the AI and non-AI recommendation type condition remained non-significant for positive emotionality ratings, p = .428, and interpersonal impression ratings, p = .854.

Table A5. *Descriptive Values for Dependent Variables in Baseline-analysis.*

	Message Type		
	Unrevised	Original AI	Original Non-AI
	M (*SD*)	*M* (*SD*)	*M* (*SD*)
Negative Emotionality Recommended			
Negative Emotionality	3.73 (1.19)	3.73 (1.31)	3.72 (1.29)
Concession Intentions	5.19 (2.70)	5.20 (2.44)	5.16 (2.56)
Positive Emotionality Recommended			
Positive Emotionality	4.01 (1.85)	3.82 (1.93)	3.72 (1.79)
Interpersonal Impression	4.10 (2.23)	3.78 (2.30)	3.75 (2.29)

Description of the Employed AI-Based Writing Assistant

The development of this tool was specifically motivated by its application in our employed experimental paradigms. Hence, it is designed to facilitate recommendations for the emotionality of input messages that are likely to appear in a psychological negotiation paradigm while allowing for the experimental manipulation of whether a positive, negative, or neutral emotionality is recommended and whether or not recommendations highlight words in an input message that are relevant to a given emotionality. It uses a *classifier* to determine the likelihood of an input message (i.e. a participant's message) having a positive, negative or neutral emotionality and an *explainer* to return the most important words for a given classification. The training data for the classifier was collected in an experimental paradigm to allow insight into its characteristics and forego 'Blackbox'-effect of commercial closed-source writing assistants.

Data Collection

Participants & Design

The training data (angry, happy, and neutral negotiation messages) was collected in a scenario study modelled as closely as possible after the negotiation paradigm in which we planned to employ the finished AI with the aim to specifically collect messages that are likely to arise within the negotiation paradigm. The study had a within-subjects design with three conditions for the negotiation course described to participants (anger vs. happiness vs. no strong emotion-inducing) to collect messages with different emotional tonalities from each participant and therefore increase the chance of parallel messages. With these measures, we aimed to increase the likelihood of our AI performing well despite a comparatively small training data set.

A sample of *N* = 300 adult native English speakers was recruited via Prolific. The recruited participants had a mean age of 35.1 years (range = 18–79), 34.0% identified as male, 65.7% identified as female, and 95.3% were UK nationals. All participants gave informed consent, were given the option to retract their data at the end of the study (which took about 9 minutes on average) and were compensated with £1.50 each. Overall, the study cost £600, including £150 in service fees.

Procedure

Each participant was presented with three different versions of a scenario in which they took the role of a customer negotiating the terms of a bicycle repair with a shop via written messages. Each version described the same negotiation course up to the shop's last offer, which was changed in each version to elicit either anger, happiness, or no strong emotion in participants. The versions' order of presentation was randomized across participants, and their efficacy in eliciting the respectively desired emotion was successfully pre-tested in a previous study. After reading each scenario, participants were asked to enter a message expressing how they felt about their negotiation partner's offers up to this point ('How do you feel about your negotiation partner's offers up to this point?') in a free text entry field. They were instructed to write exactly one sentence and to phrase it the same way they would a message to a real-life negotiation partner. This yielded 900 sentences (300 happy, angry, and non-emotional messages each) to be submitted to preprocessing.

Data Sets

We created two data sets from the initially collected 900 sentences, that were then combined for model training and testing. For the first data set, the *annotated data set*, the collected messages were annotated by three independent raters (the data set was split between raters so that each message was rated by two raters). Raters were instructed in writing with a guideline document to annotate a) the emotional content ('Is an emotion implicitly or explicitly communicated?') b) the source of the communicated emotion ('What does the communicated emotion relate to?'), and c) the type of emotion ('Which emotion is communicated?'). The possible annotation labels per annotation category are reported in Table A6. Each message could only receive one label per category from each rater.

Table A6. *Annotation Categories and Corresponding Labels.*

	Emotional Content	Emotion Source	Emotion Type
1	Emotion explicitly mentioned	Related to offer/negotiation	Happiness, Satisfaction
2	Emotion implicitly communicated	Related to negotiation partner	Anger, Frustration, Irritation
3	No emotion communicated	Related to own person	Sadness, Desperation, Disappointment, Worry
4	—	Ambiguous/Unclear	Guilt, Embarrassment
5	—	Not applicable	Ambiguous/More than one emotion
6	—	—	Not applicable

Raters agreed moderately on the explicitness of emotion communication (agreement on 75.7% of messages, Cohen's $\kappa = .60$) and on the source of the communicated emotions (agreement on 70% of messages, Cohen's $\kappa = .40$). For the most crucial annotation category, emotion type, we initially observed moderate agreement, as raters agreed on 64.5% of messages (Cohen's $\kappa = .54$). For further use in model training, the messages were categorized into positive (emotion type label 1), negative (emotion type label 2, 3, and 4), and neutral emotionality (emotion type label 5 and 6). Annotators' agreement for these summarized labels was substantial (agreement on 76.4% of messages, Cohen's $\kappa = .64$). The *annotated data set* contains 900 messages (labels are 46.3% negative,

38.4% positive, 15.2% neutral) with an average sentence length of 79.34 characters (incl. space characters; averages per label: negative = 87.53, positive = 65.63, neutral = 89.15) and 1,149 unique terms (converted to lowercase, excl. punctuation; unique terms per label: negative = 877, positive = 549, neutral = 484).

To create the second data set, the *translated data set*, a subsample of the annotated data set was manually 'translated' between the three emotional tonalities (positive vs. negative vs. neutral) by the first author and student assistants. The subsampling process was based on message quality according to annotation: A message was selected if both raters agreed that it a) either explicitly or implicitly communicated an emotion, b) communicated the negotiation/offers as the source of the emotion, and c) communicated either positive (emotion type annotation 1) or clearly negative (emotion type annotation 2 or 3) emotions. The selected messages' emotionality was then categorized as positive or negative based on annotations and messages were translated into the respective other two emotional tonalities with as few changes to content as possible. The translated data set contains 1,218 messages (labels are 33.3% positive, 33.4% negative, 33.2% neutral) with an average sentence length of 76.86 characters (incl. space characters; averages per label: negative = 78.90, positive = 75.92, neutral = 75.77) and 818 unique terms (converted to lowercase, excl. punctuation; unique terms per label: negative = 716, positive = 670, neutral = 656).

Finally, the annotated data set and the translated data set were combined into a training data split (1,151 messages), a validation data split (306 messages), and a testing data split (266 messages). In the aggregation process, the neutral classes in the annotated data and the translated data were aggregated into two separate categories due to a preprocessing mistake. Therefore, the classifier has been trained to distinguish positive and negative categories from two separate neutral representations. No categories of different semantics have been confounded and therefore this has no impact on the applicability of the classifier. The data split characteristics are reported in Table A7.

Table A7. *Data Split Characteristics.*

Data Split		Label Distribution (%)	Average Sentence Length (Char.)	Unique Terms (Excl. Punctuation)
Training	Positive	31.5	73.44	629
	Negative	37.0	82.19	835
	$Neutral_A$	24.1	76.40	553
	$Neutral_B$	7.3	86.65	362
	Overall	—	78.36	1,021
Validation	Positive	32.7	68.25	330
	Negative	37.9	81.22	413
	$Neutral_A$	19.9	73.16	242
	$Neutral_B$	9.5	90.45	209
	Overall	—	76.25	570
Testing	Positive	31.6	78.50	318
	Negative	36.5	83.98	370
	$Neutral_A$	22.9	77.92	256
	$Neutral_B$	9.0	95.88	176
	Overall	—	81.93	503

Setup

To facilitate our experimental manipulation, we used an emotion classifier to identify the emotionality of participants' original message and an explainer to be able to highlight the most relevant words for a given emotionality (i.e., classification). We implemented our assistant on a server[5] and accessed it via a restful web service from Qualtrics (the survey tool our experiment ran in).

For the *classifier*, the training data split was tokenized with RoBERTaTokenizer and used to fine-tune RoBERTa (Liu et al., 2019), a transformer-based language model pre-trained on a large corpus of English data. The classifier returns the probability of an input message expressing each of the predicted emotion labels. The results of testing the classifier's performance with the test data split are reported in Table A8. An additional 5-fold analysis showed

a mean weighted F1-value of 0.84 (*SD* = 0.02). We concluded that the classifier could reliably classify the emotionality of typical sentences from the psychological paradigm we planned to employ.

Since the computational basis behind the classifier's decisions is not readily understandable to humans, we then implemented an *explainer* to 'verbalize' why a decision is made. We chose LIME (Ribeiro et al., 2016) because it is based on a sound and robust theory, meaning it works well with different language models. It calculates an importance score for each word in an input message (i.e., participants' original messages), indicating how much each word affects the message's classification as a given emotionality. It then returns the five most important words for the classified emotionality and their scores.

Table A8. *Results of Performance Testing of the Classifier With Test Data Split.*

Emotionality	Precision	Recall	F_1
Positive	0.90	0.87	0.88
Negative	0.85	0.91	0.88
Neutral_A	0.79	0.80	0.80
Neutral_B	0.52	0.41	0.46

In the current study, the classifier's output was used to determine the most probable emotionality of participants' original message and the assistant's recommended emotionality was adapted accordingly so that a different emotionality than the one that was already expressed was recommended. In the detailed AI-based recommendation type conditions, the explainer's output was used to highlight the most important word for the undesired emotionality expressed in participants' original messages (i.e., the word with the highest weight for the emotionality classification) by highlighting the word with the highest importance score.

References

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://doi.org/10.48550/ARXIV.1907.11692

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

About Authors

Josephine Hagedorn received her PhD in Psychology from the University of Tübingen (Germany) for her research examining the impact of AI-based writing-recommendations on interpersonal online interactions. She completed this work while working as a researcher at the Leibniz-Institut für Wissensmedien (IWM), Tübingen (Germany).

https://orcid.org/0009-0006-0868-2984

Roman Klinger is a professor at the Faculty for Information Systems and Applied Computer Science (WIAI) at the University of Bamberg. His goal is to enable computers to understand and generate text regarding both propositional and non-propositional information. This finds application in interdisciplinary research, including biomedical text mining, digital humanities, modelling psychological concepts (like emotions) in language, and social media mining. These topics often constitute novel challenges to existing machine learning methods which then leads to methodological advancements in artificial intelligence.

https://orcid.org/0000-0002-2014-6619

Kai Sassenberg is director of the Leibniz Institute for Psychology and full professor at Trier University (Germany). His research focuses on emotion, motivation, and self-regulation in the context of (a) social power and leadership, (b) the processing of (in-)correct information (e.g., conspiracy theories), and (c) the change of strong attitudes such as stereotypes and norms. In addition, he is interested in the acceptance of digital technologies – for instance artificial intelligence.

https://orcid.org/0000-0001-6579-8250

Correspondence to
Josephine Hagedorn, Leibniz-Institut für Wissensmedien, Schleichtstraße 6, 72076 Tübingen, Germany,
mail@jhagedorn.de