

When Machines Stir Our Conscience – How Conversational Agents Trigger Negative Moral Emotions and the Use of the Macbeth Effect for Counteraction

Janek Mücksch, Alfred Benedikt Brendel, Florian U. Siems

Business & Information Systems Engineering (2026)

Appendix (available online via <http://link.springer.com>)

Appendix A. Study 1 Details

The following introduction was presented to the participants before starting the experiment:

Dear participant,

Thank you very much for taking the time to participate in this study and contributing to current research on online support with chatbots. During the study, you will assemble a fictional burger menu, supported by a chatbot. Moreover, we will ask you some questions in a questionnaire. We want to find out to what extent chatbots can accompany orders. You are randomly assigned to one of seven groups. This survey takes about 10 minutes to complete. If you have any further queries or observations arising from this survey, please feel free to contact the researcher Janek Mücke (Janek.muecke@tu-dresden.de).

Please note: *In the following experiment, you will assemble a burger menu. The dishes will also be displayed graphically. In exceptional cases, this can be stressful for some individuals. You therefore have the option to interrupt or cancel the study at any time if you are not feeling well. If you have been diagnosed with an eating disorder or suspect that you have an eating disorder, we ask you not to participate in the survey. An eating disorder can be, for example, anorexia, bulimia, binge-eating-disorder or mixed forms of these. Any information you provide will not be used for commercial purposes.*

All information given is confidential and only for scientific purposes. Due to the anonymous nature of this study, answers usually cannot be withdrawn after the survey has been completed.

Before interacting with the chatbot, the participants were provided with the following information on a separate webpage:

In the following, imagine you are very hungry right now. Maybe you have had a hard day at work or you are back from sports. You don't feel like cooking today. Therefore, you decide to order burgers. Imagine that you have a new app installed on your device. This application contains a chatbot and is supposed to help you to eat healthy. A chatbot is a program which is designed in order to have conversations with human users, usually via the Internet. Please select one or two burgers you would like to eat. Additionally, it is up to you whether you want to select fries and a drink as well. Please use the chatbot in the right bottom corner as a support. Although this is an experiment, please behave as realistic as possible.

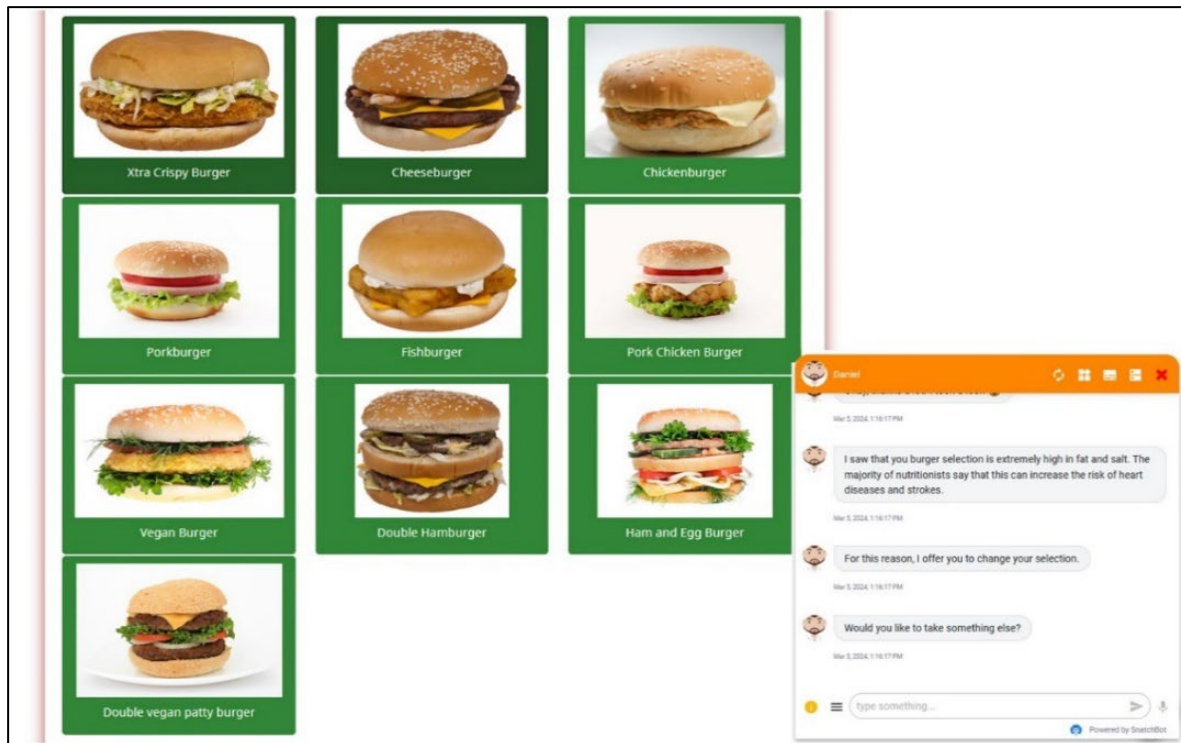


Figure A.1 Illustration of the burger selection Interface, accompanied by a humanlike CA

Table A.1 Country of residence of the respondents

Country	Amount	Share
Austria	1	0,4%
Belgium	2	0,8%
Chile	3	1,3%
Estonia	5	2,1%
France	1	0,4%
Germany	1	0,4%
Greece	5	2,1%
Hungary	7	2,9%
Israel	1	0,4%
Italy	16	6,7%
Latvia	4	1,7%
Mexico	12	5,0%
Netherlands	6	2,5%
Poland	38	15,9%
Portugal	39	16,3%
Slovenia	3	1,3%
South Africa	77	32,2%
Spain	7	2,9%
Switzerland	1	0,4%
United Kingdom	6	2,5%
United States	1	0,4%
Another country	3	1,3%

Table A.2 Measurement validation of constructs

Latent variable	Mean	SD	Loading
Moral emotions (Zhong and Liljenquist 2006) (Cronbach's α =0.905, CR =0.908, AVE =0.727)			
Shame	1.556	1.207	0.864
Guilt	1.799	1.467	0.873
Embarrassment	1.632	1.263	0.883
Regret	1.753	1.412	0.794
Disgust	1.498	1.212	0.626
Service satisfaction (Yun and Park 2022, based on Li et al. 2021) (Cronbach's α = 0.877, CR = 0.911, AVE = 0.801)			
Overall, I am satisfied with my experience of using this chatbot.	4.586	1.874	0.796
This chatbot did what I expected.	4.603	1.767	0.742
This chatbot did a good job.	4.711	1.784	0.997

Table A.3 Measurement validation of control constructs

Latent variable	Mean	SD	Loading
Social presence (Gefen and Straub 2003; Adam et al. 2021) (Cronbach's $\alpha = 0.950$, CR = 0.946, AVE = 0.792)			
There is a sense of human contact in the interaction with the chatbot.	3.732	1.859	0.885
There is a sense of personalness in the interaction with the chatbot.	3.598	1.833	0.794
There is a sense of sociability in the interaction with the chatbot.	3.728	1.828	0.889
There is a sense of human warmth in the interaction with the chatbot.	3.381	1.861	0.943
There is a sense of human sensitivity in the interaction with the chatbot.	3.356	1.797	0.929
Persuasiveness (Lehto et al. 2012) (Cronbach's $\alpha = 0.847$, CR = 0.853, AVE = 0.662)			
The chatbot has an influence on me.	2.710	1.792	0.884
The chatbot is personally relevant for me.	2.895	1.919	0.673
The chatbot makes me reconsider my eating habits.	2.798	1.975	0.872
<i>Note.</i> CR, composite reliability; M, mean; SD, standard deviation; AVE, average variance extracted.			

Table A.4 Fornell–Larcker criterion

Constructs	1.	2.	3.	4.	5.
1. Cleansing	1.000				
2. Moral emotions	-0.101	0.852			
3. Service satisfaction	-0.008	-0.088	0.858		
4. Decision adjustment	-0.050	0.238	0.049	1.000	
5. Treatment	0.035	0.422	-0.132	0.240	1.000
<i>Notes.</i> The diagonal indicates the square root of the AVE. The correlation coefficients are plotted below the diagonal. The correlations are always lower than the square root of the AVE.					

Table A.5 Descriptive statistics (Study 1)

		Cleansing		No cleansing		Two-way ANOVA
		No informing (N = 55)	Informing (N = 55)	No informing (N = 69)	Informing (N=60)	
Negative moral emotions	Mean SD	1.206 0.444	1.854 1.129	1.191 0.526	2.417 1.560	C: F(1,237) = 2.494, p = 0.115 I: F(1,237) = 50.865, p < 0.001*** C*I: F(3,235) = 4.777, p = 0.029*
Service satisfaction	Mean SD	5.085 1.424	4.157 1.707	4.648 1.614	4.666 1.662	C: F(1,237) = 0.028, p = 0.867 I: F(1,237) = 4.013, p = 0.046* C*I: F(3,235) = 5.137, p = 0.024*
Decision adjustment	Mean SD	0.236 0.429	0.382 0.490	0.217 0.415	0.517 0.504	C: F(1,237) = 0.598, p = 0.440 I: F(1,237) = 14.445, p < 0.001*** C*I: F(3,235) = 1.659, p = 0.199

Note. SD, standard deviation; C, cleansing; I, informing; * p < 0.05; ** p < 0.01; *** p < 0.001.

Table A.6 Descriptive statistics (Study 1)

		Informing		t-value	p-value (t-test)
		Cleansing (N = 55)	No cleansing (N = 60)		
Negative moral emotions	Mean SD	1.854 1.129	2.417 1.560	-2.230	p = 0.028* (two-sided) p = 0.014* (one-sided)
Service satisfaction	Mean SD	4.157 1.707	4.666 1.662	-1.619	p = 0.108 (two-sided) p = 0.054 (one-sided)
Decision adjustment	Mean SD	0.382 0.490	0.517 0.504	-1.453	p = 0.149 (two-sided) p = 0.074 (one-sided)

Note. SD, standard deviation; * p < 0.05; ** p < 0.01; *** p < 0.001.

Table A.7 Results of hypothesis tests (Study 1)

H.	Relationship	β-value	t-value	p-value (path)	Supported
H1a	Information → Moral emotions	0.542	7.889	<0.001***	Yes
H1b	Cleansing → Moral emotions	-0.232	2.153	0.016*	Yes
H1c	Cleansing*Information → Moral emotions	-0.251	2.264	0.012*	Yes
H2a	Information → Decision adjustment	0.080	2.439	0.007**	Yes
H2b	Moral emotions → Decision adjustment	0.078	2.315	0.010*	Yes
H3	Moral emotions → Service satisfaction	-0.074	1.009	0.157	No

Note. * p < 0.05; ** p < 0.01; *** p < 0.001.

For hypothesis H3, we found no support for a significant effect. Therefore, hypothesis equivalence testing was conducted (see Table A.8). More precisely, we tested whether there was no effect on a certain dependent variable. In contrast to common null hypothesis testing, we apply equivalence testing to evaluate the null hypothesis that there was an effect in these cases. An alternative hypothesis is that

there is no such effect. Contrary to common null hypothesis testing, the null hypothesis and the alternative hypothesis are reversed. (Robinson et al. 2005). Consequently, the following hypotheses are applied:

$$H_0: \beta_{Moral\ Emotions \rightarrow Service\ Satisfaction} \neq 0$$

$$H_1: \beta_{Moral\ Emotions \rightarrow Service\ Satisfaction} = 0$$

Equivalence bounds indicate a range in which no effect can be assumed (Lakens et al. 2018; Mazzolari et al. 2022). With a margin of 0.25, two one-sided t-tests were applied, both of which had to be rejected (Mazzolari et al. 2022). This results in the calculation of the following t-values:

$$t_U = \frac{\beta_{Moral\ Emotions \rightarrow Service\ Satisfaction} - \Delta_U}{SD_{Moral\ Emotions \rightarrow Service\ Satisfaction}} = \frac{-0.074 - 0.25}{0.079} = -4.101 \rightarrow p < .001$$

$$t_L = \frac{\beta_{Moral\ Emotions \rightarrow Service\ Satisfaction} - (-\Delta_L)}{SD_{Moral\ Emotions \rightarrow Service\ Satisfaction}} = \frac{-0.074 + 0.25}{0.079} = -2.228 \rightarrow p = .013$$

As this is common practice (Lakens 2017), we only report the greatest p-value and the corresponding t-value.

These equivalence tests revealed significantly no effect ($p = 0.013$; rejecting both hypotheses and the null hypothesis that there is an effect). Thus, we concluded that negative moral emotions had significantly no effect on service satisfaction.

Table A.8 Results of post-hoc tests (Study 1)

H	Relationship	β -value	p-value (path)	p-value (equivalence testing)	H0 decision (equivalence testing)
H3	Moral emotions $\rightarrow$ Service satisfaction	-0.074	0.157	0.013*	Reject
Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.					

Robustness checks:

Affinity for fast food (Dunn et al. 2011) was measured using the following items:

- Fast food tastes good
- I feel satisfied after eating fast food
- I get good value for money from fast-food meal deals
- I treat/reward myself with fast food

Reliability (Parasuraman et al. 1988; Li et al. 2021; Yun and Park 2022) was measured using the following items:

- This chatbot is dependable.
- When I have problems, this chatbot is sympathetic and reassuring.
- I felt that I could rely on the chatbot's services to fulfill my needs.

In additional robustness checks, we calculated whether applying a PROCESS model delivers the same results as a PLS-SEM. Table A.9 (PROCESS) provides the same test results as Table A.7 (PLS-SEM).

Table A.9. Results of hypothesis tests, applying PROCESS (Study 1)

H.	Relationship	β -value	t-value	p-value (path)	Supported
H1a	Information → Moral emotions	0.539	7.738	<0.001***	Yes
H1b	Cleansing → Moral emotions	-0.231	2.149	0.016*	Yes
H1c	Cleansing*Information → Moral emotions	-0.248	2.228	0.013*	Yes
H2a	Information → Decision adjustment	0.080	2.445	0.007**	Yes
H2b	Moral emotions → Decision adjustment	0.078	2.328	0.010*	Yes
H3	Moral emotions → Service satisfaction	-0.075	1.132	0.129	No
<i>Note.</i> * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.					

Appendix B. Study 2 Details

The following introduction was presented to the participants before starting the experiment:

Dear participant,

Thank you very much for taking the time to participate in this study and contributing to current research on food orders. During the study, you will assemble a fictional burger menu. Moreover, we will ask you some questions in a questionnaire. We want to find out to what extent systems can accompany orders. You are randomly assigned to one of seven groups. This survey takes about 10 minutes to complete. If you have any further queries or observations arising from this survey, please feel free to contact the researcher Janek Mücke (Janek.muecke@tu-dresden.de).

Please note: *In the following experiment, you will assemble a burger menu. The dishes will also be displayed graphically. In exceptional cases, this can be stressful for some individuals. You therefore have the option to interrupt or cancel the study at any time if you are not feeling well. If you have been diagnosed with an eating disorder or suspect that you have an eating disorder, we ask you not to participate in the survey. An eating disorder can be, for example, anorexia, bulimia, binge-eating-disorder or mixed forms of these.*

Any information you provide will not be used for commercial purposes. All information given is confidential and only for scientific purposes. Due to the anonymous nature of this study, answers usually cannot be withdrawn after the survey has been completed.

Before interacting with the point-and-click interface, the participants were provided with the following information on a separate webpage:

In the following, imagine you are very hungry right now. Maybe you have had a hard day at work or you are back from sports. You don't feel like cooking today. Therefore, you decide to order burgers. Imagine that you have a new app installed on your device. This application allows you to order food and is supposed to help you to eat healthy. Please select one or two burgers you would like to eat. Additionally, it is up to you whether you want to select fries and a drink as well.

Although this is an experiment, please behave as realistic as possible.

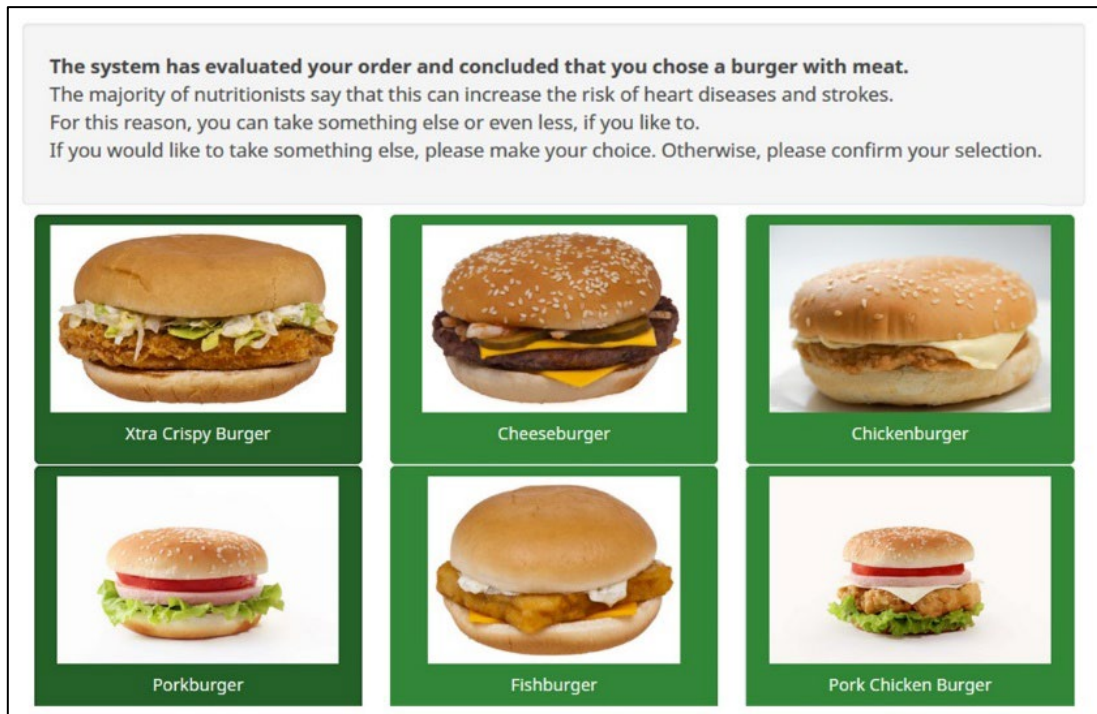


Figure B.1 Information about negative health consequences by the point-and-click interface

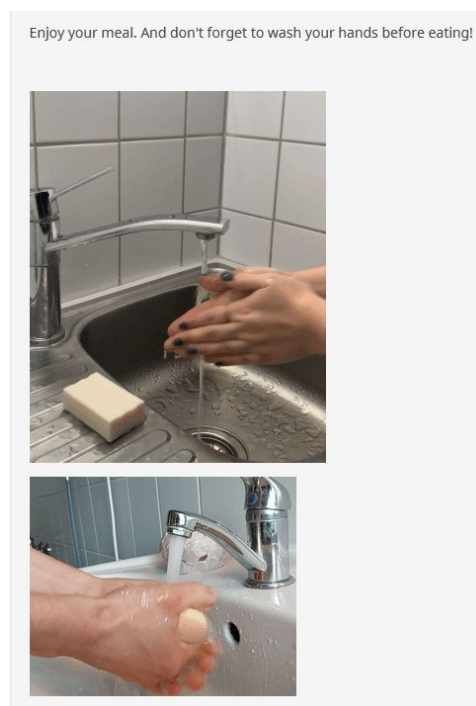


Figure B.2 Provision of the handwashing visualizations by the point-and-click interface

Table B.1 Country of residence of the respondents

Country	Amount	Share
Belgium	2	0,8%
Canada	3	1,2%
Chile	3	1,2%
Czech Republic	1	0,4%
Estonia	1	0,4%
France	4	1,6%
Greece	9	3,7%
Hungary	4	1,6%
Ireland	1	0,4%
Israel	3	1,2%
Italy	15	6,2%
Latvia	1	0,4%
Mexico	5	2,1%
Netherlands	8	3,3%
Poland	21	8,6%
Portugal	30	12,3%
South Africa	107	44,0%
Spain	6	2,5%
Sweden	1	0,4%
United Kingdom	15	6,2%
United States	1	0,4%
Another country	2	0,8%

Table B.2 Measurement validation of constructs

Latent variable	Mean	SD	Loading
Moral emotions (Zhong and Liljenquist 2006) (Cronbach's $\alpha = 0.886$, CR = 0.904, AVE = 0.682)			
Shame	1.613	1.195	0.875
Guilt	1.921	1.445	0.844
Embarrassment	1.572	1.078	0.820
Regret	1.711	1.282	0.790
Disgust	1.543	1.103	0.562
Service satisfaction (Yun and Park 2022, based on Li et al. 2021) (Cronbach's $\alpha = 0.865$, CR = 0.919, AVE = 0.787)			
Overall, I am satisfied with my experience of using this system.	5.564	1.202	0.762
This system did what I expected.	5.486	1.427	0.734
This system did a good job.	5.580	1.238	0.991
<i>Note.</i> CR, composite reliability; M, mean; SD, standard deviation; AVE, average variance extracted.			

Table B.3 Measurement validation of control constructs

Latent variable	Mean	SD	Loading
Social presence (Gefen and Straub 2003; Adam et al. 2021) (Cronbach's $\alpha = 0.950$, CR = 0.946, AVE = 0.792)			
There is a sense of human contact in the interaction with the system.	2.934	1.778	0.818
There is a sense of personalness in the interaction with the system.	3.268	1.865	0.766
There is a sense of sociability in the interaction with the system.	3.078	1.887	0.878
There is a sense of human warmth in the interaction with the system.	2.745	1.754	0.946
There is a sense of human sensitivity in the interaction with the system.	2.765	1.769	0.941
Persuasiveness (Lehto et al. 2012) (Cronbach's $\alpha = 0.789$, CR = 0.795, AVE = 0.568)			
The system has an influence on me.	2.926	1.810	0.874
The system is personally relevant for me.	3.333	1.823	0.774
The system makes me reconsider my eating habits.	3.350	1.906	0.603
<i>Note.</i> CR, composite reliability; M, mean; SD, standard deviation; AVE, average variance extracted.			

Table B.4 Fornell–Larcker criterion

Constructs	1.	2.	3.	4.	5.
1. Cleansing	1.000				
2. Moral emotions	-0.091	0.828			
3. Service satisfaction	-0.012	-0.231	0.883		
4. Decision adjustment	-0.024	-0.004	-0.006	1.000	
5. Treatment	-0.004	0.045	-0.114	0.099	1.000
<i>Notes.</i> The diagonal indicates the square root of the AVE. The correlation coefficients are plotted below the diagonal. The correlations are always lower than the square root of the AVE.					

Table B.5 Descriptive statistics (Study 2)

		Cleansing		No Cleansing		Two-Way ANOVA
		No Informing (N=61)	Informing (N=60)	No Informing (N=62)	Informing (N=60)	
Negative moral emotions	Mean	1.412	1.757	1.838	1.715	C: F(1,241) = 2.090, p = 0.150
	SD	0.658	1.108	1.161	1.100	I: F(1,241) = 0.675, p = 0.412 C*I: F(3,239) = 3.154, p = 0.077
Service satisfaction	Mean	5.844	5.207	5.538	5.598	C: F(1,241) = 0.084, p = .772
	SD	0.987	1.072	1.114	1.309	I: F(1,241) = 3.785, p = 0.053 C*I: F(3,239) = 5.797, p = 0.017*
Decision adjustment	Mean	0.300	0.433	0.361	0.419	C: F(1,241) = 0.142, p = 0.706
	SD	0.462	0.500	0.484	0.497	I: F(1,241) = 2.366, p = 0.125 C*I: F(3,239) = 0.358, p = 0.550

Note. SD, standard deviation; C, cleansing; I, informing; * p < 0.05; ** p < 0.01; *** p < 0.001.

Table B.6 Results of hypothesis tests (Study 2)

H.	Relationship	β -value	t-value	p-value (path)	Supported
H1a	Information $\rightarrow$ Moral emotions	-0.056	0.549	0.291	See Table B.7
H1b	Cleansing $\rightarrow$ Moral emotions	-0.182	1.364	0.086	
H1c	Cleansing*Information $\rightarrow$ Moral emotions	0.204	1.615	0.086	
H2a	Information $\rightarrow$ Decision adjustment	0.048	1.559	0.060	Yes
H2b	Moral emotions $\rightarrow$ Decision adjustment	-0.004	0.122	0.451	No
H3	Moral emotions $\rightarrow$ Service satisfaction	-0.231	3.920	<0.001***	Yes
Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.					

Post-Hoc Tests

As in Study 1, we tested whether there was an effect on a certain dependent variable in certain cases when we found no significant impact in the prior analysis (Table B.7).

The results indicated that informing users about the negative consequences of their choices had significantly no effect on their negative moral emotions ($p = 0.026$). Moreover, informing had significantly no impact on the Decision adjustment ($p < 0.001$), as indicated by p-values smaller than 0.05. In addition, regarding the impact and moderation of cleansing and informing, no significant effect on negative moral emotions was observed. However, it could not be confirmed that there was significantly no effect on the negative moral emotions (indicated by $p = 0.288$ and $p = 0.323$, both greater than 0.05).

Table B.7 Results of post-hoc tests (Study 2)

Relationship		β -value	p-value (path)	p-value (equivalence testing)	H0 decision (equivalence testing)
H1a	Information $\rightarrow$ Moral emotions	-0.056	0.291	0.024*	Reject
H1b	Cleansing $\rightarrow$ Moral emotions	-0.182	0.086	0.302	Do not reject
H1c	Cleansing*Information $\rightarrow$ Moral emotions	0.204	0.086	0.353	Do not reject
H2a	Information $\rightarrow$ Decision adjustment	0.048	0.060	<0.001***	Reject
H2b	Moral emotions $\rightarrow$ Decision adjustment	-0.004	0.451	<0.001***	Reject
<i>Note.</i> * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.					

Table B.8 Comparison of Study 1 and Study 2

H.	Relationship	Study 1 β-value (sd)	Study 2 β-value (sd)	Difference	t-value	p-value	Supported
H1a	Information → Moral emotions	0.542 (0.070)	-0.056 (0.098)	0.598	4.989	<0.001***	Yes
H1b	Cleansing → Moral emotions	-0.232 (0.107)	-0.182 (0.131)	-0.050	-0.295	0.385	No
H1c	Cleansing*Information → Moral emotions	-0.251 (0.111)	0.204 (0.122)	-0.455	-2.769	0.003**	Yes

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

References

- Adam M, Wessel M, Benlian A (2021) AI-based chatbots in customer service and their effects on user compliance. *Electron Mark* 31(2):427–445. <https://doi.org/10.1007/s12525-020-00414-7>
- Dunn KI, Mohr P, Wilson CJ, Wittert GA (2011) Determinants of fast-food consumption. An application of the Theory of Planned Behaviour. *Appetite* 57(2):349–357. <https://doi.org/10.1016/j.appet.2011.06.004>
- Gefen D, Straub D (2003) Managing User Trust in B2C e-Services. *e-Service Journal* 2(2):7–24. <https://doi.org/10.2979/esj.2003.2.2.7>
- Lakens D (2017) Equivalence Tests: A Practical Primer for *t* Tests, Correlations, and Meta-Analyses. *Soc Psychol Personal Sci* 8(4):355–362. <https://doi.org/10.1177/1948550617697177>
- Lakens D, Scheel AM, Isager PM (2018) Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science* 1(2):259–269. <https://doi.org/10.1177/2515245918770963>
- Lehto T, Oinas-Kukkonen H, Drozd F (2012) Factors Affecting Perceived Persuasiveness of a Behavior Change Support System. In: *International Conference on Interaction Sciences*
- Li L, Lee KY, Emokpae E, Yang S-B (2021) What makes you continuously use chatbot services? Evidence from chinese online travel agencies. *Electron Mark* 31(3):575–599. <https://doi.org/10.1007/s12525-020-00454-z>
- Mazzolari R, Porcelli S, Bishop DJ, Lakens D (2022) Myths and methodologies: The use of equivalence and non-inferiority tests for interventional studies in exercise physiology and sport science. *Exp Physiol* 107(3):201–212. <https://doi.org/10.1113/EP090171>
- Parasuraman A, Zeithaml VA, Berry LL (1988) SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *J Retail* 64(1):12–40
- Robinson AP, Duursma RA, Marshall JD (2005) A regression-based equivalence test for model validation: shifting the burden of proof. *Tree Physiol* 25(7):903–913. <https://doi.org/10.1093/treephys/25.7.903>
- Yun J, Park J (2022) The Effects of Chatbot Service Recovery With Emotion Words on Customer Satisfaction, Repurchase Intention, and Positive Word-Of-Mouth. *Front Psychol* 13:922503. <https://doi.org/10.3389/fpsyg.2022.922503>
- Zhong C-B, Liljenquist K (2006) Washing Away Your Sins: Threatened Morality and Physical Cleansing. *Science, New Series* 313(5792):1451–1452