Investigating Consumer Responses to Brand Activism and Artificial Intelligence-driven Tools

Doctoral Thesis

submitted to

Justus-Liebig Universität Gießen
Faculty of Economics and Business Studies

Giessen, 31st of July, 2023

by

Martin Haupt

Supervisors:

Prof. Dr. Alexander Haas

Chair of the Department of Marketing and Sales Management

Justus-Liebig University Giessen

Prof. Dr. Jan Freidank

Professor of International and Strategic Management

THM Business School

Acknowledgments

I cannot express enough my gratitude for all those who have supported me in this dissertation process over the last six years.

Ladies first: My heartfelt gratitude to Prof. Dr. Stefanie Wannow for your brilliant guidance, expert advice and your patience with me. It is an honor to work with you!

Katja Haupt, thank you so much for being my lighthouse in stormy times and the love of my life. I am incredibly fortunate to have you by my side!

My sincere gratitude goes to my PhD supervisors: Prof. Dr. Jan Freidank, thank you very much for your constant motivation, your strong vision, your deep understanding of the relevance of our research topics and your support in many ways! Prof. Dr. Alexander Haas, thank you very much for your open-mindedness towards various topics in marketing and for teaching me to be precise. Your feedback and review were often difficult to deal with, but have greatly advanced our research!

Finally, I would like to thank God for his guidance and blessings in my life, and my two little men, Samuel and Jonathan, for so many beautiful moments and for making me laugh countless times.

Articles in Dissertation

Brand activism

1. Wannow, S., Haupt, M. and Ohlwein, M. (2023), “Is brand activism an emotional affair? The role of moral emotions in consumer responses to brand activism”, published in: *Journal of Brand Management,* online first. https://doi.org/10.1057/s41262-023-00326-9

2. Haupt, M., Wannow, S., Marquardt, L., Graubner, J.S. and Haas, A. (2023), "Who is more responsive to brand activism? The role of consumer-brand identification and political ideology in consumer responses to activist brand messages", published in: *Journal of Product & Brand Management*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JPBM-10-2022-4193

Artificial intelligence-driven tools

3. Haupt, M., Wannow, S., Freidank, J., Haas, A., “AI wrote this? Effects of AI-powered content creation on employer attractiveness”, (working paper under review at *European Journal of Marketing*)

4. Haupt, M., Freidank, J., Haas, A., “Using Human-AI Collaborative Authorships to Escape Algorithm Aversion in Content Marketing”, (working paper in revision phase at *Review of Managerial Science*)

5. Haupt, M., Rozumowski, A., Freidank, J., Haas, A., “Seeking empathy or suggesting a solution? Effects of chatbot messages on service failure recovery to optimize digital platform efficiency” (working paper in revision phase at *Electronic Markets*)

Table of Content

List of Figures VI

List of Tables VIII

List of Appendices X

List of Abbreviations XI

General Introduction and Dissertation outline 1

I - Paper 1: Is brand activism an emotional affair? The role of moral emotions in consumer responses to brand activism 5

1. Introduction 7

2. Conceptual framework and hypotheses development 10

  2.1. Brand activism 10

  2.2. Moral emotions and consumer behaviour 12

  2.3. The role of moral emotions as a mediator between brand activism and consumer responses 15

  2.4. Consumer-brand identification as a moderator 19

3. Empirical studies 22

  3.1. Pre-studies 23

  3.2. Study 1 24

  3.3. Study 2 27

  3.4. Study 3 29

4. Discussion 31

5. References 38

II - Paper 2: Who is more responsive to brand activism? The role of consumer–brand identification and political ideology in consumer responses to activist brand messages . 62

1. Introduction 64

2. Conceptual framework 67

  2.1. CBI as a boundary condition of consumers' responses to brand activism 67

  2.2. The role of political ideology and perceived marginalization by the brand 73

  2.3. Asymmetric effects of brand activism 76

  2.4. The role of perceived marginalization as a mediator between brand activism and consumers' brand-related responses 78

3. Empirical studies 80

  3.1. Pilot study and an overview of main studies 80

3.2. Study 1 81
3.3. Study 2 85
3.4. Study 3 88
4. Discussion 90
5. References 98

III - Paper 3: AI wrote this? Effects of AI-powered content creation on employer attractiveness 124
1. Introduction 126
2. Conceptual Framework 129
2.1. AI Performance and Algorithm Aversion 129
2.2. Readers' Perceptions of AI-Created Content 130
2.3. Creating Social Media Content for Employer Branding 133
2.4. Hypotheses Development 134
4. Empirical Study 138
5. Discussion 144
6. References 151

IV - Paper 4: Using Human-AI Collaborative Authorships to Escape Algorithm Aversion in Content Marketing 162
1. Introduction 165
2. Theoretical background and hypotheses development 170
2.1 Performance and perceptions of AI 170
2.2 Human-AI collaboration as possible escape to negative reactions to AI 175
2.3 The moderating role of morality of AI use 179
3. Empirical Studies 182
3.1. Study 1 182
3.2. Study 2 185
4. Discussion 190
5. References 198

V - Paper 5: Seeking empathy or suggesting a solution? Effects of chatbot messages on service failure recovery to optimize digital platform efficiency 213
1. Introduction 215
2. Conceptual background 218
2.1. Chatbots as digital conversational agents 218

2.2. Chatbot response failures ... 218
2.3. Chatbots and the Stereotype Content Model ... 220
2.4. Recovery strategies for chatbot failure ... 222
2.5. Factors influencing the perception of recovery messages ... 227
3. Empirical studies ... 231
3.1. Pilot study ... 231
3.2. Study 1 ... 231
3.3. Study 2 – Failure frequency ... 234
3.4. Study 3 – Failure Attributions ... 238
4. Discussion ... 240
5. References ... 248

Affidavit ... 265

List of Figures

I. Is brand activism an emotional affair? The role of moral emotions in consumer responses to brand activism

Figure 1. Conceptual framework 49
Figure 2. Results from study 1 49
Figure 3. Results from study 2 50
Figure 4. Results from study 3 50

II. Who is more responsive to brand activism? The role of consumer-brand identification and political ideology in consumer responses to activist brand messages

Figure 1. Conceptual model 108
Figure 2 Estimated marginal means related to CBI and political ideology depending on consumer-brand agreement versus disagreement (Study 1) 108
Figure 3 Results of spotlight analyses (Study 2) 109
Figure 4 Results from Study 3 109

III. AI wrote this? Effects of AI-powered content creation on employer attractiveness

Figure 1. Conceptual model 155
Figure 2. Effect of Author Disclosure on Employer Attractiveness ratings 155
Figure 3. Effect of Author Expectation and Author Disclosure on Employer Attractiveness 156

List of Figures (continued)

IV. Using Human-AI Collaborative Authorships to Escape Algorithm Aversion in Content Marketing

Figure 1. Conceptual model 205
Figure 2. Study 1. Readers' perceptions of Share of Input and Level of Control 205
Figure 3. Study 1. Exemplary Scenario 206
Figure 4. Study 2. Exemplary Scenario 206
Figure 5. Study 2. Influence of Morality of AI use on relationship between authors and message credibility 207

V. Seeking empathy or suggesting a solution? Effects of chatbot messages on service failure recovery to optimize digital platform efficiency

Figure 1. Conceptual model 255
Figure 2. Exemplary scenarios 255

List of Tables

I. Is brand activism an emotional affair? The role of moral emotions in consumer responses to brand activism

Table 1. Overview of prior research regarding brand activism (BA) 51

Table 2. Scale items and statistics 53

Table 3. Study 1: Conditional process model for moral emotions as parallel mediators, CBI as moderator, and brand attitude as outcome 55

Table 4. Study 2: Conditional process model for moral emotions as parallel mediators, CBI as moderator, and brand attitude as outcome 56

Table 5. Study 3: Mediation model for moral emotions as parallel mediators and brand attitude as outcome 57

II. Who is more responsive to brand activism? The role of consumer-brand identification and political ideology in consumer responses to activist brand messages

Table 1 Scale items and statistics 110

Table 2. Repeated measures ANCOVA for brand attitude (Study 1) 111

III. AI wrote this? Effects of AI-powered content creation on employer attractiveness

Table 1. Overview of Prior Research regarding AI created Content 157

Table 2. Robustness check results comparing AI text with Human text and AI text with human as declared author 159

List of Tables (continued)

IV. Using Human-AI Collaborative Authorships to Escape Algorithm Aversion in Content Marketing

Table 1. Selected literature review regarding Human-AI Collaboration 208

Table 2. Scale items and statistics 211

Table 3. Study 2. Conditional process model for message credibility as mediator, morality of AI use as moderator, and attitude towards the company as outcome 212

V. Seeking empathy or suggesting a solution? Effects of chatbot messages on service failure recovery to optimize digital platform efficiency

Table 1. Literature Review – Chatbot Service Failure and Recovery 256

Table 2. Scenario descriptions 260

Table 3. Scale items and statistics 263

Table 4. Descriptives for Study 1, 2 and 3 264

List of Appendices

I. Is brand activism an emotional affair? The role of moral emotions in consumer responses to brand activism

Appendix 1. Scenarios used in studies 1-3 58

Appendix 2. Diagnosis of possible causes of endogeneity and techniques applied to counter them 60

II. Who is more responsive to brand activism? The role of consumer-brand identification and political ideology in consumer responses to activist brand messages

Appendix 1. Literature Review – Role of Consumer-brand identification (CBI) in Controversial or Negative Brand Events 112

Appendix 2. Literature Review – Role of Political Ideology in Controversial or Negative Brand Events 117

Appendix 3. Scenarios used in studies 1-3 120

Appendix 4. Study 3. Conditional process model with perceived marginalization as mediator and brand attitude as outcome 122

III. AI wrote this? Effects of AI-powered content creation on employer attractiveness

Appendix 1. Texts from LinkedIn – Original (human written) and AI-generated version 160

Appendix 2. Scale items and statistics 161

Appendix 3. Constructs 161

List of Abbreviations

AI	Artificial intelligence
ANOVA	Analysis of Variance
AVE	Average variance extracted
CBI	consumer-brand identification
CEO	chief executive officer
CI	confidence intervals
COV	covariance
CR	composite reliability
DIY	do-it-yourself
GPT	generative pre-trained transformer (model)
IS	information systems
U.S.	United States
WOM	word-of-mouth

General Introduction and Dissertation outline

Current political and technological trends have enormous impact on marketing strategy and consumer behavior (Bhagwat et al., 2020; Longoni & Cian, 2022; Puntoni et al., 2021). To guide marketing managers and advance theoretical and practical knowledge in these turbulent times, this dissertation focuses on two highly relevant and cutting edge issues, by investigating consumer responses to *brand activism* (paper 1 and 2) and *artificial-intelligence driven tools* (papers 3-5).

Brand activism

> *"Brand activism is perhaps the most dramatic and surprising business trend of the past decade"* (Korschun, 2021, p. 11)

Many societies around the globe experience an increasing political polarization and many consumers expect brands to engage in *brand activism*, i.e., publicly taking a stand on divisive socio-political issues (Mukherjee & Althuizen, 2020). As a response, brands increasingly advocate for controversial issues such as Black Lives Matters, gun laws or LGBTQIA-rights. Although it seems tempting for brands to actively influence public opinion and to differentiate themselves from competitors at the same time, managers need to be aware about possible backlashes from those consumers who disagree with their stance. Extremely negative consumer reactions might follow, such as people burning their Nike shoes as response to the brands engagement for Black Lives Matter. Empirical evidence has already established that brand activism is a risky strategy (Bhagwat et al., 2020) and scholars call for research to help managers understand the psychological mechanisms influencing the effects of brand activism on consumer responses. Consequently, **paper 1** of this dissertation investigates the role of moral emotions (such as anger or gratitude) as mediating factor shaping their reactions when they (dis-)agree with the company's stance. Thereby, we focus on both company- and issue related responses. Moreover, **paper 2** examines the role of consumer-brand identification (CBI) and political ideology related to consumers' responsiveness to brand activism. In addition, we assess perceived marginalization as further relevant mediating factor, which explains consumer reactions in case of their disagreement with the brand's position.

Artificial intelligence (AI)

"AI is the defining technology of our times." Satya Nadella, CEO of Microsoft

As further mega-trend in marketing, the proliferation of *Artificial intelligence* (AI) *driven tools* is strongly transforming marketing activities and customer experiences (Longoni & Cian, 2022). Both scholars and practitioners acknowledge the immense and often even disruptive potential of AI-infused applications such as self-driving cars, precise customer-screening and demand-forecasting tools, or service robots (Davenport et al., 2020; Osburg et al., 2022). As one of the most pervasive and prevalent examples, the release of the AI-driven content-generation tool ChatGPT has triggered a real hype. In just five days, it has attracted more than 1 million users, making it the fastest diffusion of a new technology ever recorded (Anderson & Rainie, 2023). Several research studies (including our study in paper 3 which relies on the predecessor model of ChatGPT) revealed that AI-generated content is hardly discernible from human-authored content. Given this high performance and expected efficiency gains for marketing automation, managers are increasingly tempted to use AI as an autonomous content creator. However, the understanding of consumer expectations and responses to AI as a content author remain limited in the marketing domain. Therefore, using the example of a highly relevant marketing text related to talent attraction, **paper 3** of this dissertation compares a human-authored text with an AI-generated text. Our research investigates potentials of AI-authored texts for branding activities and explores readers' reactions to AI disclosure. Moreover, the impact of matched or violated expectations on the company's image as an employer and the role of feelings of betrayal as a mediating variable are examined.

Studies in various contexts and the results of paper 3 revealed that people tend to have an algorithm aversion, leading to negative effects when AI is disclosed. However, as transparency is going to be a mandatory legislative requirement (e.g., as regulated in the European AI Act (European Parliament, 2023)), managers are increasingly faced with the question how to use AI without risking negative consumer reactions. As a possible solution, **paper 4** investigates whether human-AI collaboration could serve as an escape from consumers' algorithm aversion. Furthermore, this research examines the effectiveness of two distinct collaboration forms (i.e., "AI supporting a human author" vs. "AI author controlled by a human"), and the moderating impact of people's (general) morality perceptions of a company's AI use.

ChatGPT and similar tools could also be integrated as digital conversational agents to fully automate various consumer-firm interactions and service processes. Despite an increasing prevalence and high potential for efficiency gains, these chatbots still often fail and recovery strategies are urgently needed. Consequently, **paper 5** evaluates the effectiveness of two prominent failure recovery messages to maintain consumer satisfaction and loyalty. In addition, effects of situational factors such as different failure attributions or a double failure are taken into account.

References (for general introduction)

Anderson, J., & Rainie, L. (2023). *Closing thoughts on ChatGPT and other steps in the evolution of humans, digital tools and systems by 2035*. PEW research center. https://www.pewresearch.org/internet/2023/06/21/closing-thoughts-on-chatgpt-and-other-steps-in-the-evolution-of-humans-digital-tools-and-systems-by-2035/

Bhagwat, Y., Warren, N. L., Beck, J. T., & Watson, G. F. (2020). Corporate sociopolitical activism and firm value. *Journal of Marketing*, *84*(5), 1–21. https://doi.org/10.1177/0022242920937000

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, *48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

European Parliament. (2023). *EU AI Act: first regulation on artificial intelligence.* European Parliament. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Korschun, D. (2021). Brand Activism Is Here to Stay: Here's Why. *NIM Marketing Intelligence Review*, *13*(2), 10–17. https://doi.org/10.2478/nimmir-2021-0011

Longoni, C., & Cian, L. (2022). Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The "Word-of-Machine" Effect. *Journal of Marketing*, *86*(1), 91–108. https://doi.org/10.1177/0022242920957347

Mukherjee, S., & Althuizen, N. (2020). Brand activism: Does courting controversy help or hurt a brand? *International Journal of Research in Marketing*, *37*(4), 772–788. https://doi.org/10.1016/j.ijresmar.2020.02.008

Osburg, V.-S., Yoganathan, V., Kunz, W. H., & Tarba, S. (2022). Can (A)I Give You a Ride? Development and Validation of the CRUISE Framework for Autonomous Vehicle Services. *Journal of Service Research*, 109467052211182. https://doi.org/10.1177/10946705221118233

Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and Artificial Intelligence: An Experiential Perspective. *Journal of Marketing*, *85*(1), 131–151. https://doi.org/10.1177/0022242920953847

Paper 1

Is brand activism an emotional affair? The role of moral emotions in consumer responses to brand activism

Status: published at the *Journal of Brand Management*
Co-Authors: Stefanie Wannow and Martin Ohlwein
Own Share: 33%

A conceptual article linked to this paper was published as:

Wannow, S. and Haupt, M. (2022), "When Brands Take a Stand: Navigating Emotional Reactions to Brand Activism", Marketing Review St. Gallen, Vol. 39 No.2, pp. 44–51.

Further empirical research connected with this paper was presented at the following referreed conference: Esposito, E., Wannow, S. and Haupt, M., "Grateful and elevated or angry and contemptuous? The role of moral emotions in consumer responses to brand activism", European Marketing Conference, Mai 2022, Budapest, Hungary.

Is brand activism an emotional affair? The role of moral emotions in consumer responses to brand activism

Abstract

As brand activism continues to move up the corporate agenda, a more comprehensive understanding of its effects is needed. This paper contributes to the evolving research by investigating moral emotions (i.e., other-praising and other-condemning emotions) as mediating factors and consumer-brand identification (CBI) as a moderator that shape consumer reactions to brand activism. Three scenario-based experiments on two divisive topics show that activist messages elicit moral emotions that determine how individuals respond to them, depending on whether or not they agree with the brand's stance. Moreover, this effect of (dis)agreeing with the brand's stance on brand attitude is moderated by CBI. In case of a strong identification, an activist message does not affect brand attitude as CBI attenuates the activation of moral emotions – both in the positive case of agreement and in the negative case of disagreement. Finally, brand activism may counter the brand's social goals, as it disproportionally motivates opponents of the brand's stand to advocate their own contrary views on the contentious issue. In sum, these findings underscore both the emotional nature of consumer reactions to brand activism and the high level of social responsibility of companies that position themselves as political actors.

Keywords: Brand activism, Brand attitude, Consumer-brand identification, Issue advocacy, Moral emotions

1. Introduction

In today’s increasingly polarised world, companies and their brands have become political actors (Korschun, Martin, & Vadakkepatt, 2020). As a consequence, a growing number of businesses are willing to speak out on morally charged issues such as abortion rights and discrimination (Marketing Week, 2022). This development seems in line with consumer demands, as consumers who feel strongly about a controversial issue expect brands to take a clear stance on it and, ideally, even lead the way in social and political change (Mukherjee & Althuizen, 2020; Swaminathan, Sorescu, Steenkamp, O’Guinn, & Schmitt, 2020; Vredenburg, Kapitan, Spry, & Kemper, 2020). Therefore, it is tempting for brands that want to stay relevant to take a stand on divisive socio-political issues (Koch, 2020). With engaging in a polarised public debate, however, a brand expresses its values that are only shared by certain consumers (Hydock, Paharia, & Blair, 2020; Wannow & Haupt, 2022). Accordingly, companies who engage in brand activism might be praised by consumers supporting the brand’s stance while receiving backlash from others.

Research has already begun to investigate the positive and negative effects of brand activism on consumer responses (e.g., Mukherjee & Althuizen, 2020; Schmidt, Ind, Guzmán, & Kennedy, 2021) and has found it to be a risky strategy as opponents may exhibit disproportionately negative reactions (Jungblut & Johnen, 2021; Mukherjee & Althuizen, 2020). However, the psychological mechanisms mediating and moderating the effects of brand activism on consumer responses, such as brand attitudes, have been identified as a research gap (Mukherjee & Althuizen, 2020; Swaminathan et al., 2020). Surprisingly, emotions have only been considered marginally so far, although they are a key variable in consumer behaviour (Bagozzi, Gopinath, & Nyer, 1999; Bruno, Melnyk, & Murray, 2022) and strong emotional responses to brand activism are a reality (Vredenburg et al., 2020).

This research gap is addressed by our first and main research objective. Based on the observation that more and more companies are making political statements and thus contributing to morally charged public debates, we seek to provide insights into emotional reactions triggered by brand activism. We propose and show that moral emotions are crucial mediators between brand activism and consumer responses. Moral emotions typically occur whenever (relevant) moral principles are violated or upheld (Haidt, 2003), as is usually the case when brands make activist statements (Sibai, Mimoun, & Boukis, 2021). Previous studies have shown that the moral values promoted by a brand can directly trigger moral emotions such as contempt, anger, and disgust (Haidt, 2007; Tangney, Stuewig, & Mashek, 2007; Xie, Bagozzi, & Grønhaug, 2015). However, the role of both positive and negative moral emotions in consumer responses to activist statements by brands has not yet been investigated.

The second and related objective results from the consensus that brand activism should not be motivated (only) by selfish goals, such as bolstering brand attitude, but by a higher purpose, i.e., creating awareness for social issues and, ultimately, social change (Hajdas & Kłeczek, 2021). Despite the widely acknowledged purpose-driven nature of brand activism (Koch, 2020; Vredenburg et al., 2020), research to date has focused on brand- or company-related outcomes. To narrow this gap, the model considers an individual's willingness to advocate for the themed issue as a social outcome of brand activism. Consumers' support for a particular issue can be linked to moral emotions since they compel individuals to act (Haidt, 2003). Moral emotions are thus directly related to an individual's behaviour (Frijda, Kuipers, & ter Schure, 1989; Lazarus, 1991; Xie et al., 2015). Therefore, in addition to our focal consumer response, brand attitude, it will be assessed whether brand activism and moral emotions lead to issue advocacy.

Further contributing to important boundary conditions of brand activism's impact, the third research objective is to examine the extent to which consumer-brand identification (CBI) moderates the effects of brand activism on consumers' emotional and attitudinal reactions. This area of research is also considered to be incomplete to date (Mukherjee & Althuizen, 2020). In the context of brand transgressions, service failures, and the like, the moderating effects of existing consumer-brand ties have already been confirmed (R. Ahluwalia, Burnkrant, &

Unnava, 2000; Antonetti & Anesa, 2017; Khamitov, Grégoire, & Suri, 2020). In the realm of activist brands, however, it has not yet been investigated whether existing bonds between consumers and brands strengthen or weaken the effects of brand activism on brand-related responses.

Based on three experiments, this paper makes several important contributions. First, positive and negative moral emotions are introduced as mediators of the effect of brand activism on brand attitude. Second, it is examined whether CBI moderates the effects of brand activism. Finally, we investigate the extent to which brand activism leads to a multiplier effect by encouraging individuals to engage with the controversial issue.

Our findings provide managers with insight into the potential consequences of brand activism for both the brand and society by revealing the underlying mediating mechanism of moral emotions. Brand activism, which by definition relates to morally charged issues, triggers both other-praising and other-condemning emotions that lead to specific consumer reactions. For example, if brands honestly advocate a position shared by the majority of their target group, mainly positive moral emotions such as gratitude are expected to strengthen the brand and increase proactive advocacy behaviour by consumers. In contrast, when most (potential) customers disapprove the brand's stance, the brand will be target of negative moral emotions, i.e., anger, contempt, and disgust, with detrimental effects on attitude toward the brand. Next to this, CBI has been found to buffer the negative consequences of disagreements between consumers and brands. Thus, a strongly identified customer base can provide a safety cushion. In summary, these findings guide managers in their challenging decision of whether and how to engage in socio-political controversies.

2. Conceptual framework and hypotheses development

2.1. Brand activism

On a macro-level, the rise of brand activism can be understood as a consequence of increasingly polarised political debates worldwide. In the United States, a deep division over socio-political issues and a polarised discourse has been diagnosed (Grubbs, Warmke, Tosi, & James, 2020; Ketron, Kwaramba, & Williams, 2022). In Europe, mainstream parties are on the decline with populist parties gaining traction in several countries, e.g., Italy and Hungary, indicating political polarization and conflicting socio-cultural values (Vachudova, 2021). Political polarization shapes diverse consumption-related behaviours (e.g., Fernandes, 2020; Jung & Mittal, 2020) including attitudes and actions towards activist brands (Ketron et al., 2022; e.g., Matos, Vinuales, & Sheinin, 2017; Mukherjee & Althuizen, 2020). On a micro-level, more and more companies and their brands are becoming political actors by actively engaging in controversial socio-political debates in an attempt to influence consumers (Weber et al., 2021).

However, it has to be noted, that there is no consensus in management and marketing science on the fundamental question of whether it is the task of companies to engage in the field of corporate social responsibility (CSR) or even brand activism (Ferrell, 2022). Already CSR is challenged by calling for a separation of government and business (S. Ahluwalia, 2022) and questioning whether business leaders are qualified to influence public policy decisions (Cronin & Kerr, 2022). If a company acts contrary to the aspirations of society, laws and public pressure from the government should make the company comply (Gaski, 2022). In contrast, proponents of respective corporate action argue that responsibility for societal issues should be a shared responsibility rather than one to be strictly divided between agents (Demuijnck & Murphy, 2022). This joint responsibility should serve as a bridge between the company and external stakeholders (Martin & Burpee, 2022). Apart from this theoretical debate, brand activism has been obviously on the rise in the last years (Marketing Week, 2022; Moorman, 2020).

Brand activism is “a purpose- and values-driven strategy in which a brand adopts a non-neutral stance on institutionally contested socio-political issues, to create social change and marketing success” (Vredenburg et al., 2020, p. 446). Academics consider it an evolution of CSR (Sarkar & Kotler, 2018). While CSR activities are widely supported in society, scholars accentuate the contentious, often polarizing nature of the issues addressed through brand activism as a differentiator. Importantly, authentic brand activism is rooted in both corporate values and practices; otherwise consumers might accuse the brand of speaking up solely for egoistic motives and, thus, of woke washing (Mirzaei, Wilkie, & Siuki, 2022; Vredenburg et al., 2020). Still, typical topics underlying brand activism are often not linked to the core offering of the company (Vredenburg et al., 2020).

Due to its controversial nature, brand activism is considered a risky strategy, as it can potentially alienate stakeholders (Bhagwat, Warren, Beck, & Watson, 2020). With more consumers demanding brands to engage in socio-political debates, companies must weigh the risk of either taking the ’wrong‘ side on a controversial issue versus not taking a position at all (Bhagwat et al., 2020; Korschun, Aggarwal, & Rafieian, 2016; Swaminathan et al., 2020). Furthermore, engaging in a polarised public debate could limit the brand’s strategic flexibility. Public opinion is frequently quite dynamic, but a brand risks being locked into a particular stance to maintain its authenticity and credibility (Sibai et al., 2021). Apart from the potential downsides for the brand, it is unclear whether or under which circumstances brand activism contributes to a further polarization of society. There is surely a certain risk – at least for individuals with extreme political ideologies – that merely addressing a contested political issue can reinforce polarization processes regardless of the stance taken due to biased information processing (Bliuc, Bouguettaya, & Felise, 2021; Wojcieszak, Azrout, & Vreese, 2018; Wojcieszak & Warner, 2020).

Although still at an early stage, research on brand activism already offers some valuable insights, especially on brand-related outcomes and their antecedents (see Table 1 for an overview). Several studies confirm the alignment between the consumer's and the brand's stance as a key predictor of responses to brand activism, such as purchase intention (Dodd & Supa, 2014; Jungblut & Johnen, 2021; Mukherjee & Althuizen, 2020). However, studies have identified a negative net effect of brand activism on brand attitude and purchase intention (Mukherjee & Althuizen, 2020), market share (Hydock et al., 2020), and firm value (Bhagwat et al., 2020). Social impact, ultimately one of the key goals of brand activism, on the other hand, has not been the focus of empirical studies. Knowledge about the role of mediating factors also seems to be patchy. Exceptions to this are the constructs identification (Hydock et al., 2020; Mukherjee & Althuizen, 2020) and perceived hypocrisy (Korschun et al., 2016).

Research suggests that the overall impact of brand activism depends on various constraints. For instance, brand activism has been found to be riskier for larger companies as they are likely to lose more customers than they gain (Hydock et al., 2020). Furthermore, the company's value orientation matters. Individuals expect value-oriented (vs. results-oriented) companies in particular to take a stand; otherwise, they would be perceived as hypocritical (Korschun et al., 2016). Accordingly, the relevance of brand identity congruence, authenticity, and credibility in driving brand equity and social change has been emphasised (Bhagwat et al., 2020; Eilert & Nappier Cherup, 2020; Mirzaei et al., 2022; Sibai et al., 2021; Vredenburg et al., 2020). As with alternative (especially societal) outcomes of brand activism and additional mediators, there is a research gap regarding potential moderators. In the following, we will first introduce moral emotions as relevant mediating factors and link them to consumer responses, before discussing CBI as a moderating factor.

2.2. Moral emotions and consumer behaviour

Moral judgment and decision-making have long been understood from a predominantly cognitive perspective relying on the assumption of reasoned, deliberate thought processes (Haidt, 2001, 2003; Vélez García & Ostrosky-Solís, 2006; Xie et al., 2015). In the domain of brand activism research, Mukherjee and Althuizen (2020) explicitly assume a deliberate moral reasoning process to explain consumer reactions to brand activism. However, today, the idea that individuals make intuitive – often emotionally charged – moral judgements, is established as a complementary theoretical perspective (Brescoll, Okimoto, & Vial, 2018; Cova, Deonna, & Sander, 2015; Fernandes, 2020; Xie et al., 2015; Zollo, 2021). The relevance of moral intuitions was initially proposed by Haidt (2001). According to his social intuitionist model "moral judgment is caused by quick moral intuitions and is followed (when needed) by slow, ex post facto moral reasoning" (Haidt, 2001, p. 817). The (biased) reasoning why something is considered moral or immoral often only serves to justify and explain the automatic reaction (Haidt, 2001; Mercier & Sperber, 2011; Xie et al., 2015). These intuitive responses to morally significant stimuli often manifest themselves as moral emotions (Grappi, Romani, & Bagozzi, 2013b; Zollo, 2021). Additionally, Zollo (2021) argues that moral emotions also play a crucial role in the more systematic and conscious moral reasoning process that can follow the first moral intuition.

Moral emotions emerge when a specific pattern of appraisal occurs, i.e., witnessing reprehensible or exemplary moral expressions or actions that affect others (Grappi, Romani, & Bagozzi, 2013a; Haidt, 2003). Haidt (2003, p. 853) describes these emotions as 'disinterested', meaning that they are "linked to the interests or welfare either of society as a whole or at least of persons other than the judge or agent". Thus, whenever moral or ethical principles and stakes of third parties are involved, moral emotions arise (e.g., Romani, Grappi, & Bagozzi, 2013a; Xie, Bagozzi, & Grønhaug, 2019; Xu, Bolton, & Winterich, 2021; Zollo, 2021). The relevance of morality concerns in triggering emotions is also included in the cognitive appraisal theory. In this dominant theory of emotions, normative or moral compatibility has been identified as a relevant appraisal dimension (Frijda, 1986; Johnson & Stewart, 2005; Scherer, 2001).

Further, the experience of moral emotions can motivate individuals to act morally (Tangney et al., 2007). Compared to other (basic) emotions, such as joy, sadness, or fear, they are more likely to trigger prosocial behaviour (Haidt, 2003; Zollo, 2021). This applies to both positive and negative moral emotions. Thus, even emotions with a negative valence can elicit positive (i.e., prosocial) responses. For example, moral anger stimulates people to punish the ‘offender’, or to redress a perceived injustice (Haidt, 2003; van Doorn, Zeelenberg, & Breugelmans, 2014).

In the most common conceptualization of moral emotions, Haidt (2003) introduces four categories, namely (1) other-condemning, (2) other-praising, (3) self-conscious, and (4) other-suffering emotions. These four categories differ in terms of their emotional valence and their focus, i.e., is the emotion directed at the transgressor (self vs. other) or a victim. On the one hand, when someone other than the self has violated or upheld a moral value, (1) other-condemning or (2) other-praising emotions arise. On the other hand, when the moral act was performed by oneself, individuals experience (3) self-conscious emotions like guilt or pride. Lastly, when there is a clearly identifiable victim of a moral transgression, (4) other-suffering emotions like compassion are typically aroused (Greenbaum, Bonner, Gray, & Mawritz, 2020; Haidt, 2003; Zollo, 2021).

This paper focuses on other-directed moral emotions, i.e., (1) other-condemning and (2) other-praising emotions. The brand as the sender of the activist message is the transgressor or advocate of a moral value and will, therefore, be the focus of the triggered moral emotions. As the message might hurt or support one’s moral principles, we include both positive and negative moral emotions. Both categories are well-established in psychological research (Hutcherson & Gross, 2011; Landmann & Hess, 2017; Thomson & Siegel, 2017) and have been widely addressed in consumer research, particularly in the context of (un)ethical corporate behaviour, such as CSR (J. Kim & Park, 2020; Romani et al., 2013a), cause-related marketing (Le Vries & Duque, 2018), and corporate transgressions (Antonetti & Maklan, 2016; Grappi et al., 2013b; Romani, Grappi, & Bagozzi, 2013b; Xie et al., 2015). Other-condemning emotions can be defined as “negative feelings towards others because they have violated moral standards”

(Greenbaum et al., 2020, p. 96). They are typically represented by contempt, anger, and disgust (CAD) that can be collectively considered a hostility or resentment factor (Grappi et al., 2013a; Greenbaum et al., 2020; Xie et al., 2015). In contrast, other-praising emotions are triggered by exemplary moral behaviour and, therefore, represent the “brighter side to the moral emotions” (Haidt, 2003, p. 862). Gratitude and elevation are the most prototypical representatives of this category (Greenbaum et al., 2020; Haidt, 2003). People are not only moved by morally laudable actions, but also feel encouraged to act prosocial. That prosocial tendency may be focused on a particular benefactor, e.g., on a charitable brand. In addition, other-praising emotions are able to motivate people to become better persons and to positively contribute to society in general (Algoe & Haidt, 2009; Greenbaum et al., 2020; Tangney et al., 2007).

2.3. The role of moral emotions as a mediator between brand activism and consumer responses

In an activist context, other-condemning emotions can be expected to occur when the brand’s stance clashes with consumers’ moral standards, i.e., the brand’s stand is judged as morally questionable or even repulsive. In this regard, brand activism that contradicts the values of a consumer can be categorised as a values-related brand crisis (as opposed to a performance-related brand crisis) – at least from the opponents’ point of view. This type of crisis “involves social or ethical issues surrounding the values espoused by the brand” (Dutta & Pullig, 2011, p. 1282) and is not directly linked to the product or service offering (Dutta & Pullig, 2011; Liu, Lischka, & Kenning, 2018; Pullig, Netemeyer, & Biswas, 2006). For example, when a brand speaks out in favour of abortion rights, this might evoke anger, contempt, and even disgust among pro-life supporters. The arousal of negative (moral) emotions by negative values-related brand crises or ethical transgressions has been confirmed in several studies (Baghi & Gabrielli, 2019; e.g., Grappi et al., 2013a; Xie et al., 2015).

Effects on brand-related responses

Other-condemning emotions have been associated with various negative brand-related consumer responses, such as negative WOM, boycotting and protest behaviour (Antonetti & Maklan, 2016; Grappi et al., 2013a; Xie et al., 2015, 2019). We argue that individuals in this emotional state should develop a negative brand attitude besides these behavioural reactions. It has been established that emotions in general have spill-over effects and influence consumers' attitudes and judgments (Eggert, Steinhoff, & Witte, 2019; Palmatier, Jarvis, Bechkoff, & Kardes, 2009; Romani et al., 2013a), in particular when the emotions are directly related to the attitude object and are not just incidental (Pham, 2007). This is the case with brand activism which represents a purposeful brand action. Accordingly, other-condemning (and other-praising) emotions are clearly targeted at the brand (as opposed to joy or anxiety, for example). Grappi et al. (2013b) confirmed the mediating effect of anger (and gratitude) on the relationship between a company's offshoring decision, which is perceived as a moral issue by the public, and attitudes and behaviours toward the company. Leak, Woodham, and McNeil (2015) found that consumer's ideologically incongruent beliefs can cause him or her to experience anger-induced attitude change after a manager took a stance on same-sex marriage. They argue that the manager is directly tied to the brand or company; therefore, the experienced anger functions as new information to update attitudes towards the brand(s). The mediating role of contempt in the context of value-related negative brand publicity and corresponding brand evaluations has been confirmed by Liu et al. (2018). Thus, it can be hypothesised that consumer-brand disagreement on a particular topic will trigger other-condemning moral emotions, and these will subsequently negatively affect brand attitude:

> **H1a:** Consumer-brand disagreement has a negative effect on brand attitude, mediated by other-condemning emotions.

Conversely, individuals who approve of a brand's activism, i.e., perceive the brand's stance as consistent with their own moral standards, will experience other-praising emotions (Mukherjee & Althuizen, 2020). Consumers might feel supported, touched, or inspired by the brand advocating a controversial issue and appreciate the risk the brand takes (Algoe & Haidt, 2009). In this regard, if there is an agreement between the consumer and the brand, feelings of gratitude or elevation are likely to occur as the brand openly stands up for the consumer's moral values. As already argued above, emotions elicited by brand activism can be attributed to a clear source and have, in turn, a clear target, namely the brand. In consequence, it is likely that these positive moral emotions caused by purposeful brand actions will shift consumer brand attitudes in a positive direction (Pham, 2007; Wannow & Haupt, 2022). The relationship between other-praising emotions and brand-related consumer responses, mainly behavioural outcomes, has been confirmed in several studies (Eggert et al., 2019; J.-E. Kim & Johnson, 2013; J. Kim & Park, 2020; Palmatier et al., 2009; Romani et al., 2013a; Thomson & Siegel, 2017; Xie et al., 2015). Integrating an attitudinal outcome, Grappi et al. (2013b) show that gratitude mediates the effect of perceived corporate offshoring strategies on attitudes toward the company and word-of-mouth. More precisely, consumers felt grateful for the company's decision against offshoring which in turn led to more positive attitudes and behaviours. Based on the theoretical and empirical evidence presented, we hypothesise:

> **H1b:** Consumer-brand agreement has a positive effect on brand attitude, mediated by other-praising emotions.

Effects on issue advocacy

Campaigning on socio-political issues may not only elicit brand-related consumer responses (Bridger & Wood, 2017; Le Vries & Duque, 2018), but also prosocial behaviour, including issue advocacy (Romani et al., 2013a; Romani & Grappi, 2014). Issue advocacy captures a wide range of individual or collective contributions to a socio-political debate. These include participation in demonstrations or public debates (Matthes, Rios Morrison, & Schemer, 2010), as well as social media engagement, such as following or complaining to an opinion leader online (Ferrucci, Hopp, & Vargo, 2020; Xie et al., 2015).

There is empirical evidence that CSR-messages or -activities positively influence issue advocacy (i.e., prosocial behaviour) (Mantovani, Andrade, & Negrão, 2017). It has been shown that CSR increases consumers' donation intentions and actual behaviour via the other-praising emotion elevation (Romani & Grappi, 2014). On the other hand, negative emotions such as anger might motivate consumers to publicly defend their own stance, i.e., to advocate against the brand's position. Anger as part of the other-condemning emotions is often linked to a high level of activation (Rucker & Petty, 2004), leading people to blame the source of their anger and seek revenge or retaliation (Abdelwahab, San-Martín, & Jiménez, 2022; Romani, Grappi, Zarantonello, & Bagozzi, 2015). It represents a reaction to an "intentional goal blockage" (van Doorn et al., 2014, p. 261) which triggers a strong tendency to act, i.e., to remove this barrier and accomplish the pre-established goal. For instance, anger, and to a lesser extent enthusiasm, strengthens the tendency of individuals to sign a petition, participate in a demonstration, or engage in volunteer work (Feldman & Hart, 2016; Iyer, Schmader, & Lickel, 2007; Valentino, Brader, Groenendyk, Gregorowicz, & Hutchings, 2011). Previous findings also suggest that anger increases individuals' active resistance to brands (Feldman & Hart, 2016; Romani et al., 2015). All of this reflects expressive behaviour in the sense of issue advocacy.

When comparing positive and negative moral emotions, anger was found to be a stronger driver of prosocial actions than empathy or guilt, for instance in terms of advocacy for disadvantaged people (Montada & Schneider, 1989; van Doorn et al., 2014) or active political participation (Valentino et al., 2011; van Doorn et al., 2014). Therefore, negative moral emotions are hypothesised to have a stronger influence on activism-related outcomes such as issue advocacy than positive moral emotions:

> **H2**: Consumer-brand disagreement has a stronger effect on issue advocacy compared to agreement, mediated by other-condemning and other-praising moral emotions.

2.4. Consumer-brand identification as a moderator

The role of CBI in negative brand events

Responses to brand activism might depend on consumer-brand relationships. Previous research has focussed in particular on the question how existing attitudinal ties between consumers and brands influence the effect of negative – rather than positive – brand-related events, such as product or service failures and moral transgressions (e.g., R. Ahluwalia et al., 2000; Antonetti & Anesa, 2017; Davvetas & Diamantopoulos, 2017; Khamitov et al., 2020; Trump, 2014; Zhang, Zhang, & Sakulsinlapakorn, 2020). As consumers who are in opposition to a brand's stance might view such brand activism as a kind of transgression, it seems promising to extent this line of research to the phenomenon of brand activism.

In the present study, the pre-existing relationship between consumer and brand is captured by CBI (Millán & Diaz, 2014; Stokburger-Sauer, Ratneshwar, & Sen, 2012; Tuškej & Podnar, 2018). This construct is based on the long-established notion that consumers seek out products and brands that help them define and express their identity (Belk, 1988). CBI can be understood as "a consumer's perceived state of oneness with a brand" (Stokburger-Sauer et al., 2012, p. 407) which captures the degree of connectedness between one's own and the brand's identity (Davvetas & Diamantopoulos, 2017). Accordingly, CBI increases when consumers feel that they share self-defining attributes with the brand. This ultimately leads to a sense of belonging (Lam, Ahearne, Hu, & Schillewaert, 2010; Tuškej, Golob, & Podnar, 2013).

Given the symbolic function of brands, consumers for whom the brand is a core element of their identity (high CBI) should find brand activism highly relevant to their sense of self (Escalas & Bettman, 2005; Stokburger-Sauer et al., 2012). In contrast, individuals with low

CBI do not feel deeply connected to the brand, so such brand actions are not seen as identity-relevant. Scholars have found that connected consumers process negative brand-related information in a biased way to protect their self-worth, i.e., by counter-arguing, justifying and rationalizing the misstep, or decoupling the moral judgement from its source ('love is blind-effect') (R. Ahluwalia et al., 2000; Einwiller, Fedorikhin, Johnson, & Kamins, 2006; Leak et al., 2015; Wang & Kim, 2019). For example, motivated reasoning theory explains why highly identified consumers tend to engage in defensive processing of negative brand information to uphold favourable brand beliefs, while weakly identified consumers are more motivated to form accurate judgments (Einwiller et al., 2006).

For some types of transgressions, however, the opposite effect was observed ('love becomes hate-effect') (e.g., Antonetti & Anesa, 2017; Einwiller, Lis, Ruppel, & Sen, 2019; Grégoire & Fisher, 2006). In addition to extremely negative information or information highly relevant to the self, moral transgressions also appear to trigger this reverse effect (Ma, 2020; Trump, 2014). Since consumers were found to perceive information about (im)moral behaviour highly diagnostic (Goodwin, Piazza, & Rozin, 2014), a brand’s identity risks being severely damaged if it acts in a questionable or even reprehensible manner according to the individual’s (moral) standards (Antonetti & Anesa, 2017; Einwiller et al., 2019; Trump, 2014). Such transgression poses a significant threat to the self of consumers who strongly identify with the brand (Einwiller et al., 2019), as they use the brand as a source for shaping and portraying their selves (Johnson, Matear, & Thomson, 2011). In such a case, the negative information will be impossible to ignore (Einwiller et al., 2006). This is especially true because the perception of one's own morality is an essential component of the self-concept (Strohminger & Nichols, 2014). Consequently, consumers who strongly (vs. weakly) identify with the brand are likely to react more extremely (Bhattacharya & Sen, 2003) and to “reject more vehemently the company they no longer see as sharing their moral convictions or aspirations” (Einwiller et al., 2019, p. 4).

CBI and moral emotions

Transferring these considerations to consumer reactions in the context of brand activism, it can be assumed that a strong identification with the brand strengthens the emergence of other-directed moral emotions. Driven by a perceived threat to one's identity, which is particularly severe at a high CBI, other-condemning emotions in particular should be more pronounced (Einwiller et al., 2019; T. J. Kim, Yi, & Choi, 2020; Ma, 2020; Zhang et al., 2020). This suggestion is consistent with findings by Ma (2020) that brand crises enhance perceived anger and disappointment when consumers feel that self- and brand-defining (vs. non-defining) attributes are being undermined. In other words: Individuals should react more strongly to moral violations with other-condemning emotions when the violations threaten to cast a bad light on oneself and one's character, which is the case for highly identified consumers.

In addition to the motive of self-protection, the high expectations that connected consumers have of 'their' brands provide an additional explanation for the presumed amplification effect of CBI. Violations of these high standards lead consumers to experience pronounced negative emotions, such as betrayal and anger (Grégoire & Fisher, 2008; Ma, 2020; Zhang et al., 2020). Moreover, this reinforcing effect is supported by the cognitive appraisal theory, which posits the degree of personal or goal relevance as a key appraisal dimension (Nyer, 1997; Scherer, 1988; Watson & Spence, 2007). In general, a relevant stimulus is more likely to trigger a strong emotional reaction than an irrelevant stimulus (Nyer, 1997). Since a high CBI increases the personal relevance of brand activities, the stronger the person identifies with the brand, the stronger the emotions triggered by identity threats should be (Lam et al., 2010). Consequently, CBI can be expected to enhance the effect of disagreement with a brand's stance on other-condemning emotions:

> **H3a**: The effect of consumer-brand disagreement on other-condemning emotions is positively moderated by CBI.

Accordingly, pre-existing brand ties may amplify positive emotional responses to brand activism. Extending our line of reasoning to the positive case of consumer-brand agreement, we assume that highly identified consumers appreciate the pursuit of a socio-political or moral goal not only for its own sake, but also to verify their own moral self (Romani et al., 2013a). For these consumers (vs. consumers with low CBI), it provides an additional benefit if the brand is viewed as morally upright or ethical (Gao & Mattila, 2016). For consumers who share a close relationship with the brand, it should also be easier to view the brand as a moral exemplar and, thus, experience feelings of elevation (Thomson & Siegel, 2017). More generally, a high degree of perceived 'oneness' with the brand makes its commitment to shared values and purposes more self-relevant, which in turn evokes stronger positive emotions (Nyer, 1997).

As empirical evidence on the effect of CBI in the context of positively perceived moral acts is scarce, we relate to adjacent contexts. For instance, in the field of donations, a positive link between identification and gratitude has been established (Kwak & Kwon, 2016). Furthermore, a meta-analysis confirmed that identification has a positive impact on the effectiveness of cause-related marketing as well as the mediating role of positive (and negative) moral emotions (Fan, Deng, Qian, & Dong, 2020), but without analysing a possible interaction. In total, the theoretical arguments lead to the following hypothesis:

> **H3b**: The effect of consumer-brand agreement on other-praising emotions is positively moderated by CBI.

3. Empirical studies

Overview

Figure 1 provides an overview of the conceptual framework. We conducted three experimental studies to test the proposed hypotheses. Study 1 provides initial evidence for the mediating effect of moral emotions on the relationship between consumers' brand activism agreement and their brand attitude (H1a, H1b). In addition, this first study analyses the moderating impact of CBI on the link between consumer-brand (dis)agreement and moral emotions (H3a, H3b). Whereas studies 1 and 3 are based on the controversial topic of illegal immigration, study 2 uses the heavily contested topic of abortion rights to replicate the effects from study 1. Study 3 extends the mediating role of moral emotions to a society-related outcome, namely issue advocacy (H2).

[Insert Figure 1 about here]

3.1. Pre-studies

The aim of the pre-studies was to identify a socio-political issue suitable for brand activism and brands that can take an equally credible stand for both sides. To this end, a multi-method approach was applied. Secondary statistics were used to identify, on the one hand, topics that polarise the population (Appels, Edinger-Schons, Korschun, & Balk, 2020) and, on the other hand, brands whose offerings target as large a proportion of the population as possible. The topic search also included recommendations in existing studies (e.g., Moorman, 2020; Mukherjee & Althuizen, 2020; Vredenburg et al., 2020). The attribution of brand activism to a real-life brand was intended to ensure that respondents perceived the scenarios as vividly as possible. This resulted in a longlist of 8 issues and 27 brands.

In a second step, a shortlist for each of these was created based on two qualitative interviews with experts in the field. This led to the two issues of illegal immigrants and abortion rights and narrowed the list of brands down to 10. To prioritise them, 80 Amazon MTurk panellists were surveyed. Participants were randomly assigned to one of four scenarios (pro-immigrants, anti-immigrants, pro-life, pro-choice (see Appendix 1 for details)) and asked to what extent they considered it realistic for the respective brand to hold the opinion portrayed in the scenario.

Five brands did not meet the requirement that both alternative stances on an issue must be equally realistic for the brand. Of the remaining brands, a hypermarket chain and a home improvement retailer were selected. For these two companies – compared to the three alternative brands – the respondents considered it most realistic for the brand to take the position outlined in the scenario.

3.2. Study 1

Objective and study design

Study 1 pursued two main objectives. First, to examine whether other-condemning and other-praising emotions mediate the effect of brand activism on brand attitude. Second, to investigate whether CBI moderates this mediation. Based on the pre-study, illegal immigrants were selected as the controversial issue and the hypermarket chain was chosen as the source of the brand activist statement.

Participants were randomly assigned to one of three scenarios (between-subjects design, see Appendix 1). In these, the brand either took a ‘pro-immigrants’ (i.e., illegal immigrants may stay) or a ‘contra-immigrants’ position (i.e., illegal immigrants have to leave). The third scenario (control group) included a non-activist PR-statement about the general culture of the company. In order to verify that the other-directed emotional responses to brand activism were not the result of a general, possibly diffuse emotional state, ‘negative issue-related emotions’ were also included in the model. We controlled for pre-brand attitude and respondent’s involvement with the controversial issue.

The framework introduced by Hill, Johnson, Greco, O’Boyle, and Walter (2021) was used to evaluate possible causes of an endogeneity problem (see Appendix 2). As it is important to address the causes of endogeneity both theoretically (problem-avoiding) and empirically (problem-identifying) (Antonakis, Bendahan, Jacquart, & Lalive, 2010; Ketokivi & McIntosh, 2017), the techniques used in the studies focussed on making better choices in study design and analysis. Overall, the risk of endogeneity remaining despite all precautions is assessed as low.

Sample characteristics. Empirical analyses were based on data from 179 Amazon MTurk panellists (36.3% female; $M_{age} = 36.93$, $SD_{age} = 9.78$) who completed the online questionnaire for a small monetary compensation. To qualify, respondents had to have a human intelligence task (HIT) approval of 95% or higher and pass an attention check. On a continuum from 1 = 'very liberal' to 7 = 'very conservative', 39.1% identified themselves as (rather) liberal and 47.5% as (rather) conservative, with the remainder indicating they were politically neutral ($M = 4.15$; $SD = 2.07$).

Measures

A three-way split of the sample served as the categorical independent variable: individuals who agreed with the brand's stance, those who disagreed, and the control group. To identify the first two groups, participants were asked to indicate the extent to which they approved the statement "all illegal immigrants should be asked to leave the USA irrespective of how long they have been here" (1 = 'fully disagree', 7 = 'fully agree'). In the case of the anti-immigrant scenario, individuals who scored 3 or less were classified as disagreeing, while participants who scored 5 or more were classified as agreeing. 13 respondents with a score of 4 (neutral) were excluded from the analyses. In the case of the pro-immigrant scenario, the classification was mirrored.

Established scales were used to measure the constructs other-condemning emotions (CAD) (Xie et al., 2015), gratitude (Xie et al., 2015) and elevation (Xie et al., 2019) representing other-praising emotions (Xie et al., 2019), sadness and fear representing negative issue-related emotions (Izard, 1977), pre- (covariate) and post-experimental (outcome) brand attitude (Mukherjee & Althuizen, 2020), and CBI (Stokburger-Sauer et al., 2012). Brand attitude was measured with a seven-point semantic differential, while the other variables were measured with a seven-point Likert scale (1 = 'fully disagree', 7 = 'fully agree'). Individual's involvement with the issue was captured with a scale from Boninger, Krosnick, and Berent (1995). Each construct exhibited sufficient internal consistency as well as discriminant and convergent validity (Hulland, Baumgartner, & Smith, 2018) (see Table 2).

Results

Model 8 from PROCESS (Hayes, 2021) was applied to test the moderated mediation model (see Figure 2) including pre-brand attitude and issue involvement as covariates. Indirect effects were tested for statistical significance using 95% confidence intervals (based on 5,000 bootstrap samples).

As hypothesised (H1a), consumer-brand disagreement (vs. control) increased other-condemning emotions ($b = 4.50$, $t(170)= 5.18$, $p < .001$), whereas (H1b) consumer-brand agreement increased other-praising emotions ($b = 2.52$, $t(170)= 3.87$, $p < .001$). At the same time, neither disagreement nor agreement triggered negative issue-related emotions (each $p >$ fgenera.1). The moral emotions, in turn, influenced brand attitude, in each case in the expected direction: other-condemning emotions had a negative ($b = -.16$, $t(170)= -2.63$, $p < .01$), other-praising emotions a strong and positive effect on the dependent variable ($b = .52$, $t(170)= 7.69$, $p < .001$). Brand attitude was directly influenced by consumer-brand disagreement ($b = -1.52$, $t(170)= -2.55$, $p < .05$), but not by agreement ($p = .54$). Thus, H1a and H1b are supported (see data details in Table 3).

Regarding H3a, CBI moderated the effect of consumer-brand disagreement on brand attitude via other-condemning emotions (index of moderated mediation = .10, 95% CI [.02, .20]). For low and medium CBI levels [i.e., $M - 1SD = 2.88$ and $M = 4.68$], the indirect effects were significant, whereas for a high CBI level [i.e., $M + 1SD = 6.47$], the effect became insignificant (CBI_{low}: $b = -.44$, 95%CI [-.81; -.12], $\text{CBI}_{\text{medium}}$: $b = -.26$, 95%CI [-.50; -.07], CBI_{high}: $b = -.09$, 95%CI [-.30; .07]).

Similarly, CBI moderated the effect of consumer-brand agreement on brand attitude via other-praising emotions (index of moderated mediation = -.24, 95% CI [-.49, -.07]). More precisely, for a low level of CBI, the moderating effect was significant. Yet, for medium and high levels of CBI, the indirect effects became insignificant (CBI_{low}: $b = .63$, 95%CI [.19; 1.20], CBI_{medium}: $b = .20$, 95%CI [-.04; .43], CBI_{high}: $b = -.23$, 95%CI [-.69; .05]).

A high level of CBI seems to buffer the occurrence of moral emotions, whether condemning or praising, and to immunise the brand against brand activism. Conversely, the less individuals identify with a brand, the greater the leverage to influence their brand attitude through brand activism. As we found a buffering instead of the hypothesised reinforcing effect, H3a and H3b are not supported. When excluding the covariate issue involvement from the model, there are only minor differences with regard to the estimated parameter values, and the significance tests lead to the same result in each case. For reasons of parsimony, this covariate was therefore not included in studies 2 and 3.

[Insert Figure 2 about here]

3.3. Study 2

Objective and study design

Study 2 aimed to examine whether the findings of study 1 can be replicated in a different context. For this purpose, the controversial topic chosen this time was abortion rights, and the home-improvement retailer served as the activist brand. Participants were randomly assigned to one of three scenarios (between-subjects design). The brand took either a pro-life (i.e., against abortion) or a pro-choice stance (i.e., in favour of abortion rights) or, in a no-stance scenario, explicitly refused to take a stance. The first two scenarios were adopted from Mukherjee and Althuizen (2020) (see Appendix 1).

Sample characteristics

244 participants from MTurk (36.1% female; $M_{age} = 37.37$, $SD_{age} = 9.85$) completed the online questionnaire. As manipulation check, respondents had to correctly classify whether the brand was pro-choice or pro-life based on the scenario they had just read. Rated from 1 = 'very liberal' to 7 = 'very conservative', 44.7% identified themselves as (rather) liberal and 41.4% as (rather) conservative, with the remainder indicating they were politically neutral ($M = 3.84$; $SD = 2.19$).

Measures

The constructs consumer-brand agreement, other-condemning emotions, other-praising emotions, brand attitude, and CBI were measured with the same scales used in study 1. Since, according to the findings of study 1, gratitude and elevation are highly correlated, for reasons of parsimony, other-praising emotions were represented by gratitude in study 2. Each construct proved both valid and reliable (see Table 2).

Results

The moderated mediation model (see Figure 3) proposed was measured using PROCESS model 8 (Hayes, 2021), with pre-brand attitude serving as a covariate. Significance of the indirect effects were determined by 95% confidence intervals (based on 5,000 bootstrap samples).

Overall, the results confirm that the effects discovered in study 1 are robust to an alternative issue, an alternative brand, and the inclusion of a deliberate no stance-statement (see data in Table 4). Consumer-brand disagreement (vs. no-stance) increased other-condemning emotions ($b = 2.86$, $t(227)= 3.93$, $p < .001$). Likewise, consumer-brand agreement led to stronger other-praising emotions ($b = 1.13$, $t(227)= 1.97$, $p < .05$)). In turn, the two mediators influenced brand attitude (other-condemning emotions: $b = -.08$, $t(227)= -2.11$, $p < .05$; other-praising emotions: $b = .36$, $t(227)= 7.82$, $p < .001$).

Regarding H3a and H3b, CBI moderated the effect of consumer-brand disagreement on brand attitude via other-condemning emotions (index of moderated mediation = .03, 95% CI

[.001, .07]). For low and medium values of CBI [i.e., $M - 1SD = 2.23$ and $M = 4.25$], the indirect effect via other-condemning emotions was significant but became insignificant for high values of CBI [$M + 1SD = 6.26$] (CBI_{low}: $b = -.16$, 95%CI [-.32; -.01], $\text{CBI}_{\text{medium}}$: $b = -.09$, 95%CI [-.20; -.01], CBI_{high}: $b = -.02$, 95%CI [-.12; .05]). In contrast, CBI did not moderate the effect of consumer-brand agreement via other-praising emotions (index of moderated mediation = -.01, 95%CI [-.11, 10]). Consequently, the buffering effect of CBI found in study 1 can be replicated for the negative case of disagreement but not for the positive case of agreement.

[Insert Figure 3 about here]

3.4. Study 3

Objectives and study design

Study 3 pursued the objective to examine the influence of brand activism on subjects' behavioural intention to advocate for the controversial issue. For this purpose, the same controversial topic and brand as in study 1 were chosen. Subjects were randomly assigned to either the pro-immigrants or the contra-immigrants scenario (between-subjects design, see Appendix 1) already used in study 1.

Sample characteristics

Data from 205 Amazon MTurk members (45.1% female; $M_{age} = 38.86$, $SD_{age} = 11.21$) were ultimately used for the empirical analysis. To qualify, they had to have a HIT approval of 95% or higher and pass both an attention and a manipulation check. On a scale of political orientation (1 = ‘very liberal’, 7 = ‘very conservative’), 33.2% described themselves as (rather) liberal and 55.6% as (rather) conservative, while the rest identified themselves as politically neutral ($M = 4.29$; $SD = 1.97$). Because the groups narrowly passed the randomization check for age, it was included as a covariate in the model.

Measures

The constructs already introduced in the previous studies were measured with the same scales used in studies 1 and 2. In line with our hypothesis, we compared respondents who agreed to the brand's position to those who disagreed, excluding people with a neutral opinion. In addition, participants rated their willingness to advocate for the issue addressed in the statement using seven items introduced by Lu and Yuan (2021). One item was removed from the issue advocacy scale due to insufficient factor loading. The constructs exhibited adequate reliability and validity measures (see Table 2).

Results

To test the hypothesised mediation, model 4 from PROCESS (Hayes, 2021) was employed with pre-brand attitude and age as covariates. Indirect effects were tested for statistical significance by 95% confidence intervals (based on 5,000 bootstrap samples).

For the dependent variable issue advocacy (H2), the results essentially confirm the mediating role of moral emotions (see data in Table 5). Consumer-brand agreement (vs. disagreement) reduced other-condemning emotions (b = -1.30, t(177)= -4.53, p < .001) and increased other-praising emotions (b = 1.08, t(177)= 4.04, p < .001). As expected, both mediators enhanced issue advocacy (other-condemning emotions: b = .17, t(177)= 3.64, p < .001; other-praising emotions: b = .25, t(177)= 4.75, p < .001). The indirect effect of agreement via other-condemning emotions was negative (b = -.23, 95% CI [-.40; -.07]), whereas the effect via other-praising emotions was positive (b = .26, 95% CI [.11; .45]). This means that both positive and negative moral emotions motivate individuals to stand up for their opinions in society (see Figure 4).

Furthermore, issue advocacy was directly influenced by consumer-brand agreement (b = -.63, t(177)= -3.22, p < .01), and the total effect was negative (b = -.59, t(177)= -3.05, p < .01).

This supports H2 stating that issue advocacy is influenced more strongly by disagreement than by agreement.

[Insert Figure 4 about here]

4. Discussion

The decision to engage in brand activism is one that a company must make in a distinct field of tension. The legitimate aim of being commercially successful and increasing value for shareholders regularly conflicts with the ambition of contributing to the welfare of society as a whole. This is all the truer as brand activism is, by definition, about taking a stance that supports a specific interest group but at the same time works against the goals of its opponents. To what extent this kind of partisan behaviour should be the task of a company and contributes to the consensual solution of a fundamental societal challenge remains largely open. Experience suggests that creating frontlines does not facilitate this endeavour (Bliuc et al., 2021; Iyengar & Westwood, 2015; Levendusky, 2018; Wojcieszak et al., 2018; Wojcieszak & Warner, 2020).

Nevertheless, more and more companies are getting involved in brand activism. For them, the challenge is to optimise its impact for the brand and to achieve the goal of contributing to the solution of societal challenges. To help answer these questions, this research investigated how consumers react to brand activism, focusing on the mediating role of moral emotions and the moderating impact of CBI. Three scenario-based experiments identified the crucial role of other-praising and other-condemning emotions as mediators between consumer-brand (dis)agreement and brand attitude (studies 1 and 2) and issue advocacy (study 3), respectively. Furthermore, CBI was identified as a moderator (studies 1 and 2). Contrary to our hypothesis, consumers with a low (vs. high) level of CBI experienced stronger moral emotions in response to brand activism. No significant effect of consumer-brand (dis)agreement on brand attitude could be observed for consumers with a high level of CBI. Lastly, moral emotions have been shown to not only lead to brand-related outcomes, but also to trigger socially relevant behaviour (study 3). Interestingly, brand activism seems to motivate opponents of the brand's stance to advocate for their position more strongly than advocates.

Theoretical contributions

This work contributes to research in several ways. First, moral emotions are introduced as important factors mediating between brand activism and its effects. Brand activism elicits both positive and negative moral emotions, depending on whether the consumer's stance is consistent with or contrary to that of the brand. Because negative issue-related emotions do not mediate the effect of brand activism on brand attitude, the observed effects do not result from general emotional excitement triggered by the issue but relate specifically to the two categories of moral emotions. Hence, they contribute significantly to explaining consumers' responses to brand activism.

Second, this study expands knowledge of the importance of pre-existing bonds between consumers and brands. Both study 1 and 2 identify CBI as a shield against negative consumer reactions toward the brand. A single activist statement does not seem to be such a serious transgression as to severely affect brand attitude. Our findings confirm those of Leak et al. (2015), but contrast with other studies that have claimed the opposite effect for ethical, self-relevant, and controllable events (Grégoire & Fisher, 2006; Trump, 2014). One possible explanation could be that, unlike an orchestrated campaign, a single statement is not considered as a fundamental preference for a particular stance. Moreover, consumers might recognise the subjective nature of the moral values that are at the heart of brand activism. For instance, promoting residence rights for illegal immigrants does not reflect or violate a universally shared norm, but is based on different value priorities (Fernandes, 2020).

Third, the study contributes to the emerging field of research on businesses and individuals as political and social actors. Emotions triggered by brand activism motivate people to advocate for their socio-political opinions and values. This represents a desirable form of civic engagement and political participation. Yet, there is also a risk of a negative feedback effect (Korschun, Martin, & Vadakkepatt, 2020). The findings suggest that brand activism disproportionately activates the advocacy behaviour of individuals who want to achieve the opposite than the brand. Therefore, brand activism not only risks alienating certain consumer groups, but also that of going against its very purpose.

Managerial implications

The findings of the three studies provide marketers with important insights into whether and how they should or can effectively advocate on controversial issues. In terms of the ‘how’, the empirical findings suggest that activist messages should be designed to encourage the emergence of positive (and avoidance of negative) moral emotions. Gratitude, for instance, may be elicited if the campaign is judged to be purposeful and genuine, and if the company is credited with taking a risk in doing so (Bridger & Wood, 2017). In particular, care must be taken to ensure that the target audience perceives the company’s involvement as sincere and authentic (Le Vries & Duque, 2018); otherwise there is a risk that consumers will suspect the brand of woke washing and acting opportunistically (Mirzaei et al., 2022; Vredenburg et al., 2020; Yang & Mundel, 2021). For instance, brands should be sensitive when engaging in brand activism, and consistent in their actions and statements (i.e., having ‘moral integration’) (Sibai et al., 2021). Furthermore, the type of commitment (i.e., financial or rhetorical) and message framing need to match the brand type in order to be perceived as authentic (Ahmad, Guzmán, & Kidwell, 2022). Gratitude also plays a key role, as it serves as a motivator for (further) moral behaviour (Haidt, 2003), i.e. it evokes social engagement and reciprocity. If consumers feel grateful towards a brand, this moral emotion in turn elicits the desire to return the favour received by the brand, for example, by increased purchase or advocating intentions for this brand (Algoe & Haidt, 2009; Le Vries & Duque, 2018; Romani et al., 2013a; Xie et al., 2019). Particular caution should be taken in relation to the emotion of anger. Because activist campaigns are likely to arouse negative emotional reactions of those who disagree with the stance, brand managers must actively mitigate this risk. They should counteract the impression that the brand itself is unfair, disrespectful, or participating in a blame game in order to reduce the intensity of other-condemning emotions (Rozin, Lowery, Imada, & Haidt, 1999). If this is not successful, outraged consumers are likely to seek revenge and engage in anti-brand activism (Romani et al., 2015). Therefore, a clear own stance should always be linked to a dialogue-oriented approach that reaches out to those with a different opinion. Last but not least, building a reservoir of positive moral emotions respectively of associations that position the brand as a caretaker in the minds of its customers could generate a neutralising effect.

Since a high level of CBI provides a shield in case the brand's stance contradicts the individual's position, the respective status of the target group should be considered when planning a brand activism campaign. While enthusiastic followers of a brand are usually not deterred by brand activism contrary to their own opinion, caution should be taken with low CBI-individuals. However, (continued) brand activism against a consumer segment’s moral values threatens to reduce brand identification over time (Mukherjee & Althuizen, 2020), eroding the protective shield. Therefore, it is of importance to continuously strengthen it. This can be achieved, for example, by emphasizing consumer-brand similarities in other domains (e.g., shared quality standards and design preferences), through special relationship benefits (e.g., via loyalty programs or local community support), by highlighting brand prestige, or by creating memorable brand experiences (e.g., special events) (Stokburger-Sauer et al., 2012; Tuškej et al., 2013; Tuškej & Podnar, 2018). The relevance of a protective shield provided by a high CBI also leads to the implication that companies operating in a product market where the identification with brands is structurally low (vs. high) should take a much more cautious stance on ‘whether’ to engage in brand activism.

More generally, brand managers must weigh up the strategic relevance and weight they assign to brand activism as part of their overall brand strategy. The respective activism campaigns need to be orchestrated with the content of the regular campaigns and corporate activities. Following Sibai et al. (2021), brands need to have moral integration, i.e., they must be able to follow their moral beliefs in all situations (or 'practice what you preach'). In contrast, when consumers assume that a brand lacks moral uprightness, they devalue this brand and label it as inauthentic and opportunistic (Moorman, 2020; Sibai et al., 2021; Vredenburg et al., 2020).

However, before a brand starts to look at how its brand activism should be designed in concrete terms to achieve its goals, it should first clarify whether it wants to be a brand activist at all. Although brand activism seems to be in vogue, this question requires a conscious decision based on a comprehensive assessment of the opportunities and risks. In other words, brand activism should not be seen as a fad to follow just to be trendy. Rather, this decision calls for a strategic response that balances two objectives: the value it creates for society and the value it creates for its shareholders. As brand activism aims to promote change in society, the brand must realise that it is turning into a political actor. This reinforces its responsibility towards society. Furthermore, the potential consequences of such initiatives need to be carefully assessed in advance. These include not only the risk of losing customers and the profit they generate, but also the fact that an activism campaign encourages in particular opponents of the brand's stance to stand up for their own behalf. As a consequence, the supporters should be actively involved in the campaign and thus encouraged to promote the company's stance. Obstacles in this regard must be removed in order to mobilise (otherwise) silent supporters. In short, brand activism should not be undertaken in a perfunctory way, it should either be undertaken consistently or not at all.

Limitations and future research

Limitations of this paper provide interesting avenues for future research. First, the study focuses on the two groups of moral emotions that the individual directs toward the brand as the other element in the relationship. Including further moral emotions such as pride, shame, or guilt (self-conscious emotions) into the model might prove insightful.

Second, the scenarios presented the company's stance without providing a detailed justification of the position. Additional information or a request for understanding could be an opportunity to influence the target group's emotional response. Examining effects of more inclusive communication and de-escalation strategies on people's emotions could therefore be a task for future studies.

Third, only a single, isolated statement served as a stimulus. In reality, however, brand activism seems to be an essential part of the brand-DNA and as such a strategic decision rather than a tactical measure (Sarkar & Kotler, 2018). Longitudinal studies lend themselves to capture this. In addition, constructs such as authenticity, credibility, and consistency could provide interesting insights in this context.

Forth, further studies could examine whether activism actually serves its goal of changing society for the better. According to the findings, activism particularly benefits the opposing side's cause. Moreover, brand activism threatens to further polarise society, one of the top five challenges worldwide (World Economic Forum, 2021). Brand activism thus faces a serious dilemma. This should be elucidated in a multi-disciplinary approach by analysing in detail the long-term societal impact of brand activism.

Finally, the interplay between cognitive and affective responses to brand activism deserves further attention. The importance of cognitions and emotions in moral consumer decision-making has already been confirmed. However, the detailed hierarchy of effects and potential alternative pathways have not been assessed in the context of brand activism.

Declaration

On behalf of all authors, the corresponding author states that there is no conflict of interest.

All authors contributed equally to this work. The co-first authorship order was determined randomly.

5. References

Abdelwahab, D., San-Martín, S., & Jiménez, N. (2022). Does regional bias matter? Examining the role of regional identification, animosity, and negative emotions as drivers of brand switching: An application in the food and beverage industry. *Journal of Brand Management*, *29*, 111–126. https://doi.org/10.1057/s41262-021-00260-8

Ahluwalia, R., Burnkrant, R. E., & Unnava, H. R. (2000). Consumer response to negative publicity: The moderating role of commitment. *Journal of Marketing Research*, *37*(2), 203–214. https://doi.org/10.1509/jmkr.37.2.203.18734

Ahluwalia, S. (2022). A critique of corporate social responsibility in light of classical economics. *AMS Review*, *12*(1-2), 25–29. https://doi.org/10.1007/s13162-022-00224-4

Ahmad, F., Guzmán, F., & Kidwell, B. (2022). Effective messaging strategies to increase brand love for sociopolitical activist brands. *Journal of Business Research*, *151*, 609–622. https://doi.org/10.1016/j.jbusres.2022.07.031

Algoe, S. B., & Haidt, J. (2009). Witnessing excellence in action: The 'other-praising' emotions of elevation, gratitude, and admiration. *The Journal of Positive Psychology*, *4*(2), 105–127. https://doi.org/10.1080/17439760802650519

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*(6), 1086–1120. https://doi.org/10.1016/j.leaqua.2010.10.010

Antonetti, P., & Anesa, M. (2017). Consumer reactions to corporate tax strategies: The role of political ideology. *Journal of Business Research*, *74*, 1–10. https://doi.org/10.1016/j.jbusres.2016.12.011

Antonetti, P., & Maklan, S. (2016). An extended model of moral outrage at corporate social irresponsibility. *Journal of Business Ethics*, *135*(3), 429–444. https://doi.org/10.1007/s10551-014-2487-y

Appels, M., Edinger-Schons, L. M., Korschun, D., & Balk, L. (2020). Don't mix business with politics? Understanding stakeholder reactions to corporate political activism. *Marketing Review St. Gallen*, *37*(4), 888–894.

Baghi, I., & Gabrielli, V. (2019). The role of crisis typology and cultural belongingness in shaping consumers' negative responses towards a faulty brand. *Journal of Product & Brand Management*, *28*(5), 653–670. https://doi.org/10.1108/JPBM-03-2018-1806

Bagozzi, R. P., Gopinath, M., & Nyer, P. U. (1999). The role of emotions in marketing. *Journal of the Academy of Marketing Science*, *27*(2), 184–206. https://doi.org/10.1177/0092070399272005

Belk, R. W. (1988). Possessions and the extended self. *Journal of Consumer Research*, *15*(2), 139–168. https://doi.org/10.1086/209154

Berkowitz, L., & Harmon-Jones, E. (2004). Toward an understanding of the determinants of anger. *Emotion*, *4*(2), 107–130. https://doi.org/10.1037/1528-3542.4.2.107

Bhagwat, Y., Warren, N. L., Beck, J. T., & Watson, G. F. (2020). Corporate sociopolitical activism and firm value. *Journal of Marketing*, *84*(5), 1–21. https://doi.org/10.1177/0022242920937000

Bliuc, A.-M., Bouguettaya, A., & Felise, K. D. (2021). Online Intergroup Polarization Across Political Fault Lines: An Integrative Review. *Frontiers in Psychology*, *12*, 641215. https://doi.org/10.3389/fpsyg.2021.641215

Boninger, D. S., Krosnick, J. A., & Berent, M. K. (1995). Origins of attitude importance: Self-interest, social identification, and value relevance. *Journal of Personality and Social Psychology*, *68*(1), 61–80. https://doi.org/10.1037/0022-3514.68.1.61

Brescoll, V. L., Okimoto, T. G., & Vial, A. C. (2018). You've come a long way… maybe: How moral emotions trigger backlash against women leaders. *Journal of Social Issues*, *74*(1), 144–164. https://doi.org/10.1111/josi.12261

Bridger, E. K., & Wood, A. (2017). Gratitude mediates consumer responses to marketing communications. *European Journal of Marketing*, *51*(1), 44–64. https://doi.org/10.1108/EJM-11-2015-0810

Bruno, P., Melnyk, V., & Murray, K. B. (2022). The temperature dimension of emotions. *European Journal of Marketing*, *56*(8), 2172–2215. https://doi.org/10.1108/EJM-04-2020-0237

Cova, F., Deonna, J., & Sander, D. (2015). Introduction: Moral Emotions. *Topoi*, *34*(2), 397–400. https://doi.org/10.1007/s11245-015-9345-0

Cronin, J. J., & Kerr, J. R. (2022). Marketing's new myopia: Expanding the social responsibilities of marketing managers. *AMS Review*, *12*(1-2), 30–37. https://doi.org/10.1007/s13162-022-00228-0

Davvetas, V., & Diamantopoulos, A. (2017). "Regretting your brand-self?" The moderating role of consumer-brand identification on consumer responses to purchase regret. *Journal of Business Research*, *80*, 218–227. https://doi.org/10.1016/j.jbusres.2017.04.008

Demuijnck, G., & Murphy, P. E. (2022). Why should marketers be forced to ignore their moral awareness? A reply to Gaski. *AMS Review*, *12*(1-2), 38–43. https://doi.org/10.1007/s13162-022-00226-2

Dodd, M. D., & Supa, D. W. (2014). Conceptualizing and measuring "corporate social advocacy" communication: Examining the impact on corporate financial performance. *Public Relations Journal*, *8*(3), 2–23.

Dutta, S., & Pullig, C. (2011). Effectiveness of corporate responses to brand crises: The role of crisis type and response strategies. *Journal of Business Research*, *64*(12), 1281–1287. https://doi.org/10.1016/j.jbusres.2011.01.013

Eggert, A., Steinhoff, L., & Witte, C. (2019). Gift purchases as catalysts for strengthening customer-brand relationships. *Journal of Marketing*, *83*(5), 115–132.

Eilert, M., & Nappier Cherup, A. (2020). The activist company: Examining a company's pursuit of societal change through corporate activism using an institutional theoretical lens. *Journal of Public Policy & Marketing*, *39*(4), 461–476. https://doi.org/10.1177/0743915620947408

Einwiller, S., Fedorikhin, A., Johnson, A. R., & Kamins, M. A. (2006). Enough is enough! When identification no longer prevents negative corporate associations. *Journal of the*

Academy of Marketing Science, *34*(2), 185–194. https://doi.org/10.1177/0092070305284983

Einwiller, S., Lis, B., Ruppel, C., & Sen, S. (2019). When CSR-based identification backfires: Testing the effects of CSR-related negative publicity. *Journal of Business Research*, *104*, 1–13.

Escalas, J. E., & Bettman, J. R. (2005). Self-construal, reference groups, and brand meaning. *Journal of Consumer Research*, *32*(3), 378–389. https://doi.org/10.1086/497549

Fan, X., Deng, N., Qian, Y., & Dong, X. (2020). Factors affecting the effectiveness of cause-related marketing: A meta-analysis. *Journal of Business Ethics*, 1–22. https://doi.org/10.1007/s10551-020-04639-6

Feldman, L., & Hart, P. S. (2016). Using political efficacy messages to increase climate activism: The mediating role of emotions. *Science Communication*, *38*(1), 99–127. https://doi.org/10.1177/1075547015617941

Fernandes, D. (2020). Politics at the mall: The moral foundations of boycotts. *Journal of Public Policy & Marketing*, *39*(4), 494–513. https://doi.org/10.1177/0743915620943178

Ferrell, O. C. (2022). Perspectives on socially responsible marketing: The chasm widens. *AMS Review*, *12*(1-2), 1–6. https://doi.org/10.1007/s13162-022-00229-z

Ferrucci, P., Hopp, T., & Vargo, C. J. (2020). Civic engagement, social capital, and ideological extremity: Exploring online political engagement and political expression on Facebook. *New Media & Society*, *22*(6), 1095–1115. https://doi.org/10.1177/1461444819873110

Frijda, N. H. (1986). *The emotions*. Cambridge, UK: Cambridge University Press.

Frijda, N. H., Kuipers, P., & ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, *57*(2), 212–228. https://doi.org/10.1037/0022-3514.57.2.212

Gao, Y. L., & Mattila, A. S. (2016). The impact of option popularity, social inclusion/exclusion, and self-affirmation on consumers' propensity to choose green hotels. *Journal of Business Ethics*, *136*(3), 575–585. https://doi.org/10.1007/s10551-014-2536-6

Gaski, J. F. (2022). Toward social responsibility, not the social responsibility semblance: Marketing does not need a conscience. *AMS Review*, *12*(1-2), 7–24. https://doi.org/10.1007/s13162-022-00227-1

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research*, *44*(1), 196–210. https://doi.org/10.1093/jcr/ucx047

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Grappi, S., Romani, S., & Bagozzi, R. P. (2013a). Consumer response to corporate irresponsible behavior: Moral emotions and virtues. *Journal of Business Research*, *66*(10), 1814–1821. https://doi.org/10.1016/j.jbusres.2013.02.002

Grappi, S., Romani, S., & Bagozzi, R. P. (2013b). The effects of company offshoring strategies on consumer responses. *Journal of the Academy of Marketing Science*, *41*(6), 683–704. https://doi.org/10.1007/s11747-013-0340-y

Greco, L. M., O'Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2018). Meta-analysis of coefficient alpha: A reliability generalization study. *Journal of Management Studies*, *55*(4), 583–618. https://doi.org/10.1111/joms.12328

Greenbaum, R., Bonner, J., Gray, T., & Mawritz, M. (2020). Moral emotions: A review and research agenda for management scholarship. *Journal of Organizational Behavior*, *41*(2), 95–114. https://doi.org/10.1002/job.2367

Grégoire, Y., & Fisher, R. J. (2006). The effects of relationship quality on customer retaliation. *Marketing Letters*, *17*(1), 31–46. https://doi.org/10.1007/s11002-006-3796-4

Grégoire, Y., & Fisher, R. J. (2008). Customer betrayal and retaliation: When your best customers become your worst enemies. *Journal of the Academy of Marketing Science*, *36*(2), 247–261.

Grubbs, J. B., Warmke, B., Tosi, J., & James, A. S. (2020). Moral grandstanding and political polarization: A multi-study consideration. *Journal of Research in Personality*, *88*, 1–12. https://doi.org/10.1016/j.jrp.2020.104009

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Sherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford, UK: Oxford University Press.

Haidt, J. (2007). The new synthesis in moral psychology. *Science (New York, N.Y.)*, *316*(5827), 998–1002. https://doi.org/10.1126/science.1137651

Hajdas, M., & Kłeczek, R. (2021). The real purpose of purpose-driven branding: Consumer empowerment and social transformations. *Journal of Brand Management*, *28*(4), 359–373. https://doi.org/10.1057/s41262-021-00231-z

Hayes, A. F. (2021). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (Third edition). New York, NY: Guilford publications.

Hill, A. D., Johnson, S. G., Greco, L. M., O'Boyle, E. H., & Walter, S. L. (2021). Endogeneity: A review and agenda for the methodology-practice divide affecting micro and macro research. *Journal of Management*, *47*(1), 105–143. https://doi.org/10.1177/0149206320960533

Hulland, J., Baumgartner, H., & Smith, K. M. (2018). Marketing survey research best practices: Evidence and recommendations from a review of JAMS articles. *Journal of the Academy of Marketing Science*, *46*(1), 92–108. https://doi.org/10.1007/s11747-017-0532-y

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social-functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology*, *100*(4), 719. https://doi.org/10.1037/a0022408

Hydock, C., Paharia, N., & Blair, S. (2020). Should your brand pick a side? How market share determines the impact of corporate political advocacy. *Journal of Marketing Research*, *57*(6), 1135–1151. https://doi.org/10.1177/0022243720947682

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, *59*(3), 690–707.

Iyer, A., Schmader, T., & Lickel, B. (2007). Why individuals protest the perceived transgressions of their country: The role of anger, shame, and guilt. *Personality and Social Psychology Bulletin*, *33*(4), 572–587.

Izard, C. E. (1977). *Human Emotions*. New York, NY: Plenum Press.

Johnson, A. R., Matear, M., & Thomson, M. (2011). A coal in the heart: Self-relevance as a post-exit predictor of consumer anti-brand actions. *Journal of Consumer Research*, *38*(1), 108–125. https://doi.org/10.1086/657924

Johnson, A. R., & Stewart, D. W. (2005). A reappraisal of the role of emotion in consumer behavior: Traditional and contemporary approaches. In N. K. Malhotra (Ed.), *Review of marketing research* (Vol. 1, pp. 3–24). Bingley: Emerald Group Publishing Limited.

Jung, J., & Mittal, V. (2020). Political identity and the consumer journey: A research review. *Journal of Retailing*, *96*(1), 55–73. https://doi.org/10.1016/j.jretai.2019.09.003

Jungblut, M., & Johnen, M. (2021). When brands (don't) take my stance: The ambiguous effectiveness of political brand communication. *Communication Research*, 1-26. https://doi.org/10.1177/00936502211001622

Ketokivi, M., & McIntosh, C. N. (2017). Addressing the endogeneity dilemma in operations management research: Theoretical, empirical, and pragmatic considerations. *Journal of Operations Management*, *52*(1), 1–14. https://doi.org/10.1016/j.jom.2017.05.001

Ketron, S., Kwaramba, S., & Williams, M. (2022). The "company politics" of social stances: How conservative vs. liberal consumers respond to corporate political stance-taking. *Journal of Business Research*, *146*, 354–362. https://doi.org/10.1016/j.jbusres.2022.03.086

Khamitov, M., Grégoire, Y., & Suri, A. (2020). A systematic review of brand transgression, service failure recovery and product-harm crisis: Integration and guiding insights. *Journal of the Academy of Marketing Science*, *48*(3), 519–542. https://doi.org/10.1007/s11747-019-00679-1

Kim, J.-E., & Johnson, K. K. P. (2013). The impact of moral emotions on cause-related marketing campaigns: A cross-cultural examination. *Journal of Business Ethics*, *112*(1), 79–90. https://doi.org/10.1007/s10551-012-1233-6

Kim, J., & Park, T. (2020). How corporate social responsibility (CSR) saves a company: The role of gratitude in buffering vindictive consumer behavior from product failures. *Journal of Business Research*, *117*, 461–472. https://doi.org/10.1016/j.jbusres.2020.06.024

Kim, T. J., Yi, Y., & Choi, J. (2020). The boomerang effect of brand personality congruency in a product-harm crisis. *Australian Journal of Management*, *45*(4), 645–661. https://doi.org/10.1177/0312896219895059

Koch, C. H. (2020). Brands as activists: The Oatly case. *Journal of Brand Management*, *27*(5), 593–606. https://doi.org/10.1057/s41262-020-00199-2

Korschun, D., Aggarwal, A., & Rafieian, H. (2016). Taking a stand: Consumer responses to corporate political activism. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.2806476

Korschun, D., Martin, K. D., & Vadakkepatt, G. (2020). Marketing's role in understanding political activity. *Journal of Public Policy & Marketing*, *39*(4), 378–387. https://doi.org/10.1177/0743915620949261

Krause, M. S., & Howard, K. I. (2003). What random assignment does and does not do. *Journal of Clinical Psychology*, *59*(7), 751–766. https://doi.org/10.1002/jclp.10170

Kwak, D. H., & Kwon, Y. (2016). Can an organization's philanthropic donations encourage consumers to give? The roles of gratitude and boundary conditions. *Journal of Consumer Behaviour*, *15*(4), 348–358. https://doi.org/10.1002/cb.1576

Lam, S. K., Ahearne, M., Hu, Y., & Schillewaert, N. (2010). Resistance to brand switching when a radically new brand is introduced: A social identity theory perspective. *Journal of Marketing*, *74*(6), 128–146. https://doi.org/10.1509/jmkg.74.6.128

Landmann, H., & Hess, U. (2017). What elicits third-party anger? The effects of moral violation and others' outcome on anger and compassion. *Cognition and Emotion*, *31*(6), 1097–1111. https://doi.org/10.1080/02699931.2016.1194258

Lazarus, R. S. (1991). *Emotion and Adaptation*. Oxford, UK: Oxford University Press.

Le Vries, E. de, & Duque, L. C. (2018). Small but sincere: How firm size and gratitude determine the effectiveness of cause marketing campaigns. *Journal of Retailing*, *94*(4), 352–363. https://doi.org/10.1016/j.jretai.2018.08.002

Leak, R. L., Woodham, O. P., & McNeil, K. R. (2015). Speaking candidly: How managers' political stances affect consumers' brand attitudes. *Journal of Product & Brand Management*, *24*(5), 494–503. https://doi.org/10.1108/JPBM-08-2014-0693

Lerner, J. S., Goldberg, J. H., & Tetlock, P. E. (1998). Sober second thought: The effects of accountability, anger, and authoritarianism on attributions of responsibility. *Personality and Social Psychology Bulletin*, *24*(6), 563–574. https://doi.org/10.1177/0146167298246001

Levendusky, M. S. (2018). When Efforts to Depolarize the Electorate Fail. *Public Opinion Quarterly*, *82*(3), 583–592. https://doi.org/10.1093/poq/nfy036

Liu, X., Lischka, H. M., & Kenning, P. (2018). Asymmetric cognitive, emotional and behavioural effects of values-related and performance-related negative brand publicity. *Journal of Product & Brand Management*, *27*(2), 128–145. https://doi.org/10.1108/JPBM-11-2016-1351

Lu, H., & Yuan, S. (2021). What motivates information sharing about disaster victims on social media? Exploring the role of compassion, sadness, expectancy violation, and enjoyment. *International Journal of Disaster Risk Reduction*, *63*(2). https://doi.org/10.1016/j.ijdrr.2021.102431

Ma, L. (2020). When love becomes hate: How different consumer-brand relationships interact with crises to influence consumers' reactions. *Corporate Communications: An International Journal*, *25*(3), 357–375. https://doi.org/10.1108/CCIJ-08-2019-0103

Mantovani, D., Andrade, L. M. de, & Negrão, A. (2017). How motivations for CSR and consumer-brand social distance influence consumers to adopt pro-social behavior. *Journal of Retailing and Consumer Services*, *36*, 156–163. https://doi.org/10.1016/j.jretconser.2017.01.009

Marketing Week (2022). Authentic brands and political speech: Can companies have it both ways? Retrieved from https://www.marketingweek.com/cmo-survey-authentic-brands-political-speech/

Martin, K. D., & Burpee, S. (2022). Marketing as problem solver: In defense of social responsibility. *AMS Review*, *12*(1-2), 44–51. https://doi.org/10.1007/s13162-022-00225-3

Matos, G., Vinuales, G., & Sheinin, D. A. (2017). The power of politics in branding. *Journal of Marketing Theory and Practice*, *25*(2), 125–140. https://doi.org/10.1080/10696679.2016.1270768

Matthes, J., Rios Morrison, K., & Schemer, C. (2010). A spiral of silence for some: Attitude certainty and the expression of political minority opinions. *Communication Research*, *37*(6), 774–800. https://doi.org/10.1177/0093650210362685

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57-74. https://doi.org/10.1017/S0140525X10000968

Millán, Á., & Diaz, E. (2014). Analysis of consumers' response to brand community integration and brand identification. *Journal of Brand Management*, *21*(3), 254–272. https://doi.org/10.1057/bm.2014.4

Mirzaei, A., Wilkie, D. C., & Siuki, H. (2022). Woke brand activism authenticity or the lack of it. *Journal of Business Research*, *139*, 1–12.

Montada, L., & Schneider, A. (1989). Justice and emotional reactions to the disadvantaged. *Social Justice Research*, *3*(4), 313–344. https://doi.org/10.1007/BF01048081

Moorman, C. (2020). Commentary: Brand activism in a political world. *Journal of Public Policy & Marketing*, *39*(4), 388–392. https://doi.org/10.1177/0743915620945260

Mukherjee, S., & Althuizen, N. (2020). Brand activism: Does courting controversy help or hurt a brand? *International Journal of Research in Marketing*, *37*(4), 772–788. https://doi.org/10.1016/j.ijresmar.2020.02.008

Nyer, P. U. (1997). A study of the relationships between cognitive appraisals and consumption emotions. *Journal of the Academy of Marketing Science*, *25*(4), 296–304. https://doi.org/10.1177/0092070397254002

Palmatier, R. W., Jarvis, C. B., Bechkoff, J. R., & Kardes, F. R. (2009). The role of customer gratitude in relationship marketing. *Journal of Marketing*, *73*(5), 1–18. https://doi.org/10.1016/j.jretconser.2014.06.007

Pham, M. T. (2007). Emotion and rationality: A critical review and interpretation of empirical evidence. *Review of General Psychology*, *11*(2), 155–178. https://doi.org/10.1037/1089-2680.11.2.155

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *The Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*, 539–569. https://doi.org/10.1146/annurev-psych-120710-100452

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management*, *12*(4), 531–544. https://doi.org/10.1177/014920638601200408

Pullig, C., Netemeyer, R. G., & Biswas, A. (2006). Attitude basis, certainty, and challenge alignment: A case of negative brand publicity. *Journal of the Academy of Marketing Science*, *34*(4), 528–542. https://doi.org/10.1177/0092070306287128

Romani, S., & Grappi, S. (2014). How companies' good deeds encourage consumers to adopt pro-social behavior. *European Journal of Marketing*, *48*(5/6), 943–963. https://doi.org/10.1108/EJM-06-2012-0364

Romani, S., Grappi, S., & Bagozzi, R. P. (2013a). Explaining consumer reactions to corporate social responsibility: The role of gratitude and altruistic values. *Journal of Business Ethics*, *114*(2), 193–206. https://doi.org/10.1177/0092070306287128

Romani, S., Grappi, S., & Bagozzi, R. P. (2013b). My anger is your gain, my contempt your loss: Explaining consumer responses to corporate wrongdoing. *Psychology & Marketing*, *30*(12), 1029–1042. https://doi.org/10.1002/mar.20664

Romani, S., Grappi, S., Zarantonello, L., & Bagozzi, R. P. (2015). The revenge of the consumer! How brand moral violations lead to consumer anti-brand activism. *Journal of Brand Management*, *22*(8), 658–672. https://doi.org/10.1057/bm.2015.38

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*(4), 574. https://doi.org/10.1037//0022-3514.76.4.574

Rucker, D. D., & Petty, R. E. (2004). Emotion specificity and consumer behavior: Anger, sadness, and preference for activity. *Motivation and Emotion*, *28*(1), 3–21. https://doi.org/10.1023/B:MOEM.0000027275.95071.82

Sarkar, C., & Kotler, P. (2018). *Brand activism: From Purpose to Action* (2nd ed.). Idea Bite Press.

Scherer, K. R. (1988). Criteria for emotion-antecedent appraisal: A review. In Hamilton, V., Bower, G. H., Frijda, N. H. (Ed.), *Cognitive perspectives on emotion and motivation* (pp. 89–126). Dordrecht: Kluwer Academic/Plenum Publishers.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K.R. Scherer, A. Schorr, & T. Johnstone (Ed.), *Appraisal processes in emotion: Theory, methods, research* (pp. 92–120). New York: Oxford University Press.

Schmidt, H. J., Ind, N., Guzmán, F., & Kennedy, E. (2021). Sociopolitical activist brands. *Journal of Product & Brand Management*, *31*(1), 40-55 https://doi.org/10.1108/JPBM-03-2020-2805

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* ([Nachdr.]). Belmont, Calif.: Wadsworth Cengage Learning.

Sibai, O., Mimoun, L., & Boukis, A. (2021). Authenticating brand activism: Negotiating the boundaries of free speech to make a change. *Psychology & Marketing*, *38*(10), 1651–1669. https://doi.org/10.1002/mar.21477

Siemsen, E., Roth, A., & Oliveira, P. (2010). Common method bias in regression models with linear, quadratic, and interaction effects. *Organizational Research Methods*, *13*(3), 456–476. https://doi.org/10.1177/1094428109351241

Stokburger-Sauer, N., Ratneshwar, S., & Sen, S. (2012). Drivers of consumer-brand identification. *International Journal of Research in Marketing*, *29*(4), 406–418. https://doi.org/10.1016/j.ijresmar.2012.06.001

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005

Swaminathan, V., Sorescu, A., Steenkamp, J.-B. E., O'Guinn, T. C. G., & Schmitt, B. (2020). Branding in a hyperconnected world: Refocusing theories and rethinking boundaries. *Journal of Marketing*, *84*(2), 24–46. https://doi.org/10.1177/0022242919899905

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*, 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

Thomson, A. L., & Siegel, J. T. (2017). Elevation: A review of scholarship on a moral and other-praising emotion. *The Journal of Positive Psychology*, *12*(6), 628–638. https://doi.org/10.1080/17439760.2016.1269184

Trump, R. K. (2014). Connected consumers' responses to negative brand actions: The roles of transgression self-relevance and domain. *Journal of Business Research*, *67*(9), 1824–1830. https://doi.org/10.1016/j.jbusres.2013.12.007

Tuškej, U., Golob, U., & Podnar, K. (2013). The role of consumer–brand identification in building brand relationships. *Journal of Business Research*, *66*(1), 53–59.

Tuškej, U., & Podnar, K. (2018). Exploring selected antecedents of consumer–brand identification. *Baltic Journal of Management*, *13*(4), 451–470. https://doi.org/10.1108/BJM-08-2017-0268

Vachudova, M. A. (2021). Populism, democracy, and party system change in Europe. *Annual Review of Political Science*, *24*, 471–498. https://doi.org/10.1146/annurev-polisci-041719-102711

Valentino, N. A., Brader, T., Groenendyk, E. W., Gregorowicz, K., & Hutchings, V. L. (2011). Election night's alright for fighting: The role of emotions in political participation. *The Journal of Politics*, *73*(1), 156–170. https://doi.org/10.1017/S0022381610000939

Van Doorn, J., Zeelenberg, M., & Breugelmans, S. M. (2014). Anger and prosocial behavior. *Emotion Review*, *6*(3), 261–268. https://doi.org/10.1177/1754073914523794

Vélez García, A. E., & Ostrosky-Solís, F. (2006). From morality to moral emotions. *International Journal of Psychology*, *41*(5), 348–354. https://doi.org/10.1080/00207590500345898

Vredenburg, J., Kapitan, S., Spry, A., & Kemper, J. A. (2020). Brands taking a stand: Authentic brand activism or woke washing? *Journal of Public Policy & Marketing*, *39*(4), 444–460. https://doi.org/10.1177/0743915620947359

Wang, S., & Kim, K. J. (2019). Consumer response to negative celebrity publicity: The effects of moral reasoning strategies and fan identification. *Journal of Product & Brand Management*, *29*(1), 114–123. https://doi.org/10.1108/JPBM-10-2018-2064

Wannow, S., & Haupt, M. (2022). When Brands Take a Stand: Navigating Emotional Reactions to Brand Activism. *Marketing Review St. Gallen*. (2), 44–51.

Watson, L., & Spence, M. T. (2007). Causes and consequences of emotions on consumer behaviour: A review and integrative cognitive appraisal theory. *European Journal of Marketing*, *41*(5-6), 487–511. https://doi.org/10.1108/03090560710737570

Weber, T. J., Hydock, C., Ding, W., Gardner, M., Jacob, P., Mandel, N., . . . van Steenburg, E. (2021). Political polarization: Challenges, opportunities, and hope for consumer welfare, marketers, and public policy. *Journal of Public Policy & Marketing*, *40*(2), 184–205. https://doi.org/10.1177/0743915621991103

Wojcieszak, M., Azrout, R., & Vreese, C. de (2018). Waving the red cloth: Media coverage of a contentious issue triggers polarization. *Public Opinion Quarterly*, *82*(1), 87–109. https://doi.org/10.1093/poq/nfx040

Wojcieszak, M., & Warner, B. R. (2020). Can Interparty Contact Reduce Affective Polarization? A Systematic Test of Different Forms of Intergroup Contact. *Political Communication*, *37*(6), 789–811. https://doi.org/10.1080/10584609.2020.1760406

World Economic Forum (2021). The global risks report. Retrieved from www.weforum.org/reports/the-global-risks-report-2021

Xie, C., Bagozzi, R. P., & Grønhaug, K. (2015). The role of moral emotions and individual differences in consumer responses to corporate green and non-green actions. *Journal of the Academy of Marketing Science*, *43*(3), 333–356. https://doi.org/10.1007/s11747-014-0394-5

Xie, C., Bagozzi, R. P., & Grønhaug, K. (2019). The impact of corporate social responsibility on consumer brand advocacy: The role of moral emotions, attitudes, and individual differences. *Journal of Business Research*, *95*, 514–530. https://doi.org/10.1016/j.jbusres.2018.07.043

Xu, H. F., Bolton, L. E., & Winterich, K. P. (2021). How do consumers react to company moral transgressions? The role of power distance belief and empathy for victims. *Journal of Consumer Research*, *48*(1), 77–101. https://doi.org/10.1093/jcr/ucaa067

Yang, J., & Mundel, J. (2021). “Are We All in This Together?”: Brand Opportunism in COVID-19 Cause Related Marketing and the Moderating Role of Consumer Skepticism. *Journal of Promotion Management*, *27*(6), 877–899. https://doi.org/10.1080/10496491.2021.1888181

Zhang, Y., Zhang, J., & Sakulsinlapakorn, K. (2020). Love becomes hate? Or love is blind? Moderating effects of brand love upon consumers’ retaliation towards brand failure. *Journal of Product & Brand Management*, *30*(3), 415–432. https://doi.org/10.1108/JPBM-07-2019-2470

Zollo, L. (2021). The consumers’ emotional dog learns to persuade its rational tail: Toward a social intuitionist framework of ethical consumption. *Journal of Business Ethics*, *168*(2), 295–313. https://doi.org/10.1007/s10551-019-04420-4

Figure 5. Conceptual framework

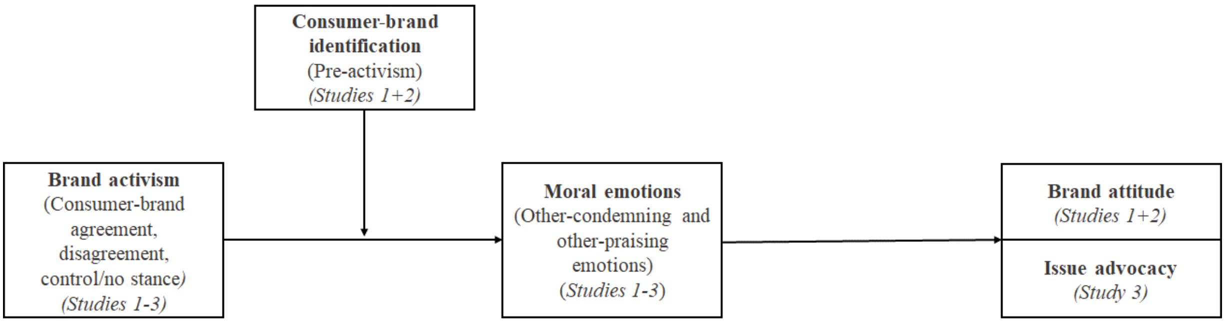
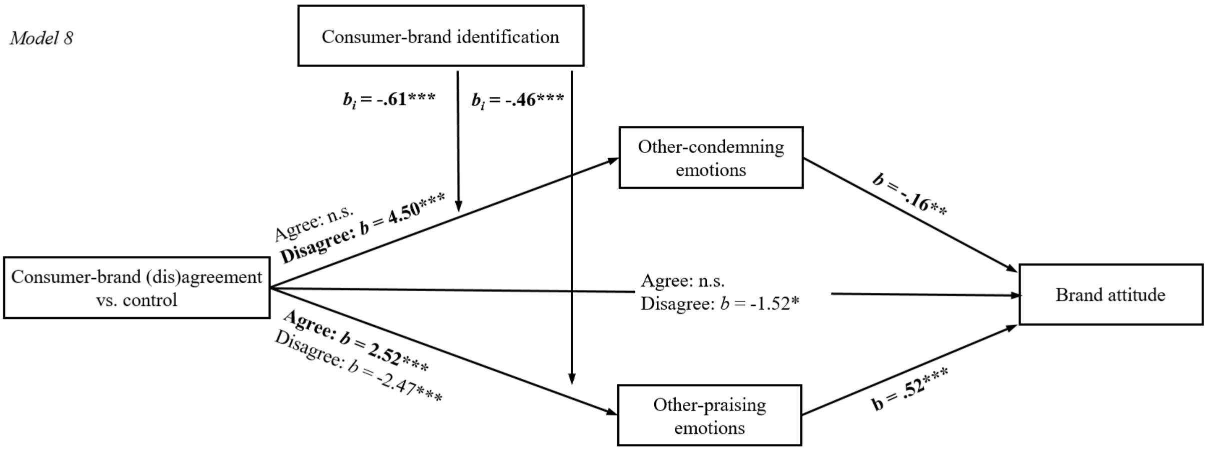
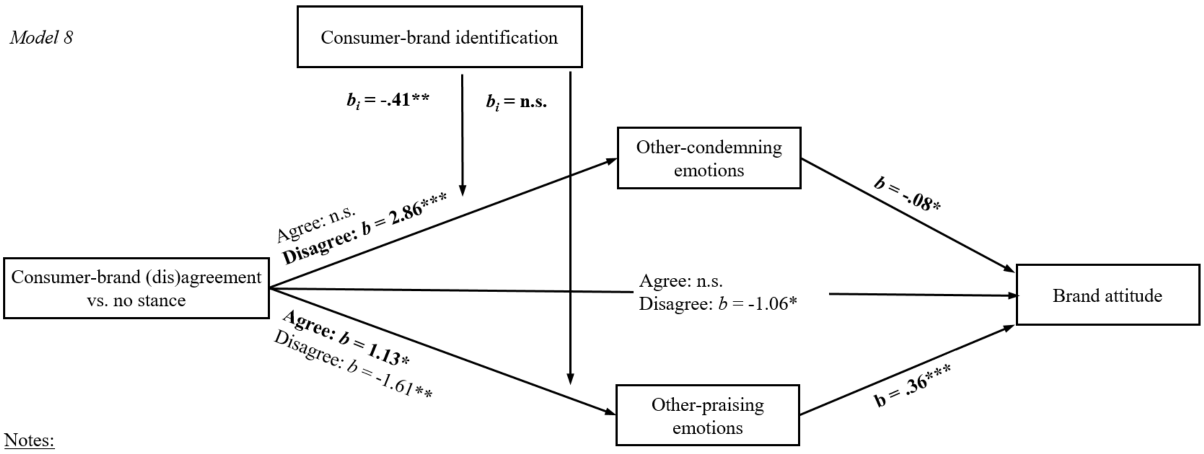


Figure 6. Results from study 1



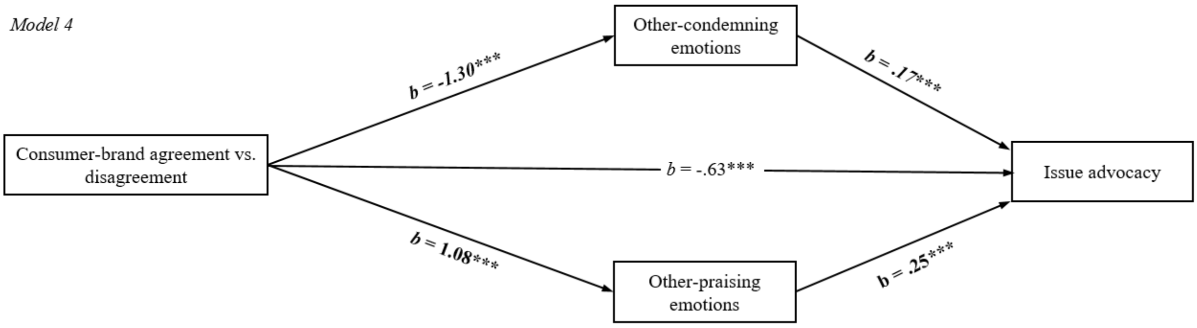
Notes:
Issue-related emotions as parallel mediator not depicted.
*** = p <.001, ** p <.01, * = p <.05, n.s. = not significant

Figure 7. Results from study 2



Notes:
*** = *p* <.001, ** *p* <.01, * = *p* <.05, n.s. = not significant

Figure 8. Results from study 3



Notes:
*** = p <.001

Table 1: Overview of prior research regarding brand activism (BA)

Study	Type	Activism issue(s)	Main finding(s)
Mirzaei, Wilkie and Siuki (2022)	Content analysis	Black lives matter, Me Too	Brand authenticity was found to be a major success factor for BA. This study identifies six dimensions for woke BA authenticity, namely social context independency, inclusion, sacrifice, practice, fit, and motivation.
Jungblut and Johnen (2021)	Experiments	LGBTQIA, gun control, immigration	BA was found to be a risky strategy for brands: When comparing buying-related responses to BA, boycotting (turning away from brand) outweighed buycotting (supporting brands). A higher political interest and a lower product category involvement decreased this imbalance between boycotting and buycotting.
Klostermann, Hydock and Decker (2021)	Quantitative	Gun control policy, immigration policy, abortion policies, etc.	BA had negative effects on brand perceptions, and this effect was stronger for customers of the brand (vs. non-customers). The effects were moderated by the level of effort and BA concurrence. Online protests were also driven by BA effort and had a negative impact on brand perception.
Schmidt et al. (2021)	Qualitative, Quantitative, Experiment	Abortion rights, LGBTQ, immigration, racial equality	Consumers increasingly expect a brand to take a stance on sociopolitical topics, but perceived brand authenticity as relevant for BA success. A brand's stance should be linked to the company's long-term commitment to the issue and be aligned with the brands values and overall strategy.
Bhagwat et al. (2020)	Quantitative	Immigration, gun control, LGBTQ, climate change	Activism leads to adverse reactions from business investors, as they perceive activism as signal towards a risky activity and away from profit-aligned targets. However, a higher match of the activism issue and stakeholder values and image reduce this negative effect. Moreover, a deliberate action vs. an activism statement reduces the negative effect of BA on shareholders' investment reactions.
Eilert and Nappier (2020)	Conceptual	Racial discrimination, transgender, immigrants, gun control	BA is framed as response to barriers which hinder changes within the society. Furthermore, effects of firm's identity orientation on BA facilitation is discussed, e.g., a social welfare identity orientation is associated with riskier BA practices.

Hydock, Paharia and Blair (2020)	Experiments	Illegal immigrants	BA is proposed to benefit small-share brands and hurt large-share brands, because large firms have a broad customer base and potentially more customers to loose than smaller ones. Authenticity is a precondition for small-share brands to benefit from BA actions.
Koch (2020)	Case Study	Climate change, food production	Food-brand Oatly is chosen to describe how BA changes consumption, production, policy, and ideologies. When brands earned legitimacy, BA can lead to brand development and positioning.
Milfeld and Flint (2020)	Qualitative	Toxic masculinity	BA videos like Gillette's 'The Best Men Can Be' can lead to alternative interpretations and lead to polarization. Using an everyday person (vs. celebrity) and subtle product placement (vs. prominent) could create higher narrative transportation (i.e., readers are more immersed into the video) and alignment with the brand's intended interpretation. High perceived realism could facilitate negative autobiographical memories.
Moorman (2020)	Conceptual	Black lives matter, same-sex marriage etc.	Seven brand political activism perspectives are proposed, which describe a brand's role regarding socio-political issues: brand authenticity, corporate citizen, cultural authority, calculative, brands as educators, political mission and employee engagement view.
Mukherjee and Althuizen (2020)	Experiments	Immigration, abortion rights, freedom of speech	BA creates asymmetric effect on brand attitude, consumers' behavioural intentions and choices. Disagreement with the brand's stance leads to negative effects, while agreement with the brand's stance has no effect. Higher distance between message sender (e.g., CEO as private person vs. official spokesperson) and brand allows moral decoupling and weakens the negative effect.
Vredenburg et al. 2020	Conceptual	Immigration, gender rights, LGBTQIA, gun laws, climate change	Brands should engage in authentic BA, i.e., matching their activist messaging with purpose, values and prosocial practice. Otherwise, consumers might perceive activism as woke washing, resulting in harm for the brand. A framework for brand-cause fit regarding BA is proposed, describing an optimal incongruence between brand and BA cause.
Leak, Woodham and McNeil (2015)	Quantitative	Same-sex marriage	People hold different and possibly conflicting ideological beliefs related to a manager's expression of their stance on a social issue. When people perceive such conflicting beliefs (vs. not) and choose to rely on an ideology that is congruent with the managers stance, this position leads to less anger and increased brand attitude. Pre-existing brand attitudes offer a halo effect and reduce the amount of anger.

Table 2. Scale items and statistics

Construct name and items	Standardised loadings		
	Study 1	Study 2	Study 3
Other-condemning emotions (CAD) (Study 1/2/3: α = .98/.97/.98; AVE = .83/.84/.83; CR = .98/.98/.98)			
• Angry	.92	.90	.88
• Mad	.88	.90	.91
• Very Annoyed	.91	.93	.88
• Scornful	.91	.93	.93
• Contemptuous	.90	.91	.91
• Disdainful	.92	.93	.92
• Feeling of Distaste	.92	.93	.93
• Disgusted	.93	.91	.92
• Feeling of Revulsion	.92	.91	.93
Other-praising emotions (Gratitude, elevation items only for Study 1) (Study 1/2/3: α = .94/.92/.96; AVE = .58/.62/.74; CR = .85/.77/.85)			
• Grateful (Gratitude 1)	.76	.81	.87
• Thankful (Gratitude 2)	.80	.77	.85
• Touched (Elevation 1)	.75		
• Moved (Elevation 2)	.74		
Consumer-brand identification (CBI) (Study 1/2/3: α = .95/.97/.95; AVE = .73/.80/.74; CR = .93/.95/.94)			
• I feel a strong sense of belonging to the brand	.89	.88	.88
• I identify strongly with the brand	.90	.91	.86
• The brand embodies what I believe in	.85	.90	.85
• The brand is like a part of me	.80	.89	.88
• The brand has a great deal of personal meaning for me	.82	.91	.87
Brand attitude (Study 1/2: α = .96/.94; AVE = .76/.64; CR = .90/.84)			
• Bad – Good	.86	.80	
• Unpleasant – Pleasant	.85	.77	
• Dislike – Like	.90	.82	
Issue advocacy (Study 3: α = .89; AVE = .56; CR = .88)			
• Express your opinion among friends, family, or colleagues			.81
• Participate in discussions in public			.74
• Take part in collections of signatures (e.g., petitions)			.67
• Like or follow an opinion leader on this issue			.76
• Vote for candidates who are committed to the issue			.80

- Donate to organizations working on the issue	.69

Harman's single factor test (Podsakoff & Organ, 1986)	Variance explained (in %) Study 1	Study 2	Study 3
Single factor	40.8	42.6	40.7

Note: All results indicate that common method bias is not an issue.

Table 3. Study 1: Conditional process model for moral emotions as parallel mediators, CBI as moderator, and brand attitude as outcome

Mediator variable models

	Other-condemning emotions		Other-praising emotions		Negative issue-related emotions	
	b	*t*	*b*	*t*	*b*	*t*
X1: Brand agreement vs. control	-.52	-.56 $^{\text{n.s.}}$	2.52	3.87 ***	-.17	-.21 $^{\text{n.s.}}$
X2: Brand disagreement vs. control	4.50	5.18 ***	-2.47	-4.04 ***	-.29	-.38 $^{\text{n.s.}}$
W: CBI	.38	2.96 **	.75	8.27 ***	.39	3.52 ***
X1*W	.14	.78 $^{\text{n.s.}}$	-.46	-3.55 ***	.11	.71 $^{\text{n.s.}}$
X2*W	-.61	-3.49 ***	.32	2.64 **	.09	.57 $^{\text{n.s.}}$
COV: Pre-brand attitude	-.16	-1.19 $^{\text{n.s.}}$	.03	.28 $^{\text{n.s.}}$	-.17	1.49 $^{\text{n.s.}}$
COV: Issue involvement	.32	3.45 ***	.07	1.03 ***	.35	4.24 ***

Outcome variable model

	Brand-attitude	
	b	*t*
X1: Brand agreement vs. control	-.34	-.61 $^{\text{n.s.}}$
X2: Brand disagreement vs. control	-1.52	-2.55 *
M: Other-condemning emotions	-.16	-2.63 **
M: Other-praising emotions	.52	7.69 ***
W: CBI	-.32	-3.33 **
X1*W	.05	.43 $^{\text{n.s.}}$
X2*W	.25	2.22 *
COV: Pre-brand attitude	.60	7.50 ***
COV: Issue involvement	-.04	-.65 $^{\text{n.s.}}$

Conditional indirect effect(s) of X on Y (brand attitude) at values of the moderator
Bootstrap 95 per cent confidence intervals for conditional indirect effect

	CBI	*b*	Lower	Upper
X1: Other-praising emotions	2.88	.63	.19	1.20
	4.68	.20	-.04	.43
	6.47	-.23	-.69	.05

	CBI	*b*	Lower	Upper
X2: Other-condemning emotions	2.88	-.44	-.81	-.12
	4.68	-.26	-.50	-.07
	6.47	-.09	-.30	.07

Note. * *p* < .05, ** *p* < .01, *** *p* < .001, n.s. = not significant.

Table 4. Study 2: Conditional process model for moral emotions as parallel mediators, CBI as moderator, and brand attitude as outcome

Mediator variable models

	Other-condemning emotions		Other-praising emotions	
	b	*t*	*b*	*t*
X1: Brand agreement vs. no stance	-.27	-.40 $^{n.s.}$	1.13	1.97 *
X2: Brand disagreement vs. no stance	2.86	3.93 ***	-1.61	-2.66 **
W: CBI	.60	4.36 ***	.29	2.52 *
X1*W	.03	.24 $^{n.s.}$	-.03	-.25 $^{n.s.}$
X2*W	-.41	-2.67 **	.30	2.39 *
COV: Pre-brand attitude	-.38	-2.84 **	.37	3.30 **

Outcome variable model

	Brand-attitude	
	b	*t*
X1: Brand agreement vs. no-stance	-.30	-.76 $^{n.s.}$
X2: Brand disagreement vs. no-stance	-1.06	-2.44 *
M: Other-condemning emotions	-.08	-2.11 *
M: Other-praising emotions	.36	7.82 ***
W: CBI	-.04	-.47 $^{n.s.}$
X1*W	.08	1.00 $^{n.s.}$
X2*W	.21	2.36 *
COV: Pre-brand attitude	.55	6.80 ***

Conditional indirect effect(s) of X on Y (brand attitude) at values of the moderator

Bootstrap 95 per cent confidence intervals for conditional indirect effect

	CBI	*b*	Lower	Upper
X1: Other-praising emotions	2.23	.39	.07	.73
	4.25	.36	.16	.61
	6.26	.34	.10	.66

	CBI	*b*	Lower	Upper
X2: Other-condemning emotions	2.23	-.16	-.32	-.01
	4.25	-.09	-.20	-.01
	6.26	-.02	-.12	.05

Note. * $p < .05$, ** $p < .01$, *** $p < .001$, n.s. = not significant.

Table 5. Study 3: Mediation model for moral emotions as parallel mediators and brand attitude as outcome

Mediator variable models

	Other-condemning emotions		Other-praising emotions	
	b	*t*	*b*	*t*
X: Brand agreement vs. disagreement	-1.30	-4.53 ***	1.08	4.04 ***
COV: Pre-brand attitude	.04	.37 $^{n.s.}$	.44	4.30 ***
COV: Age	-.00	-.77 $^{n.s.}$	-.02	-1.61 $^{n.s.}$

Outcome variable model

	Issue Advocacy	
	b	*t*
X: Brand agreement vs. disagreement	-.63	-3.22 **
M: Other-condemning emotions	.17	3.64 ***
M: Other-praising emotions	.25	4.75 ***
COV: Pre-brand attitude	.04	.53 $^{n.s.}$
COV: Age	-.01	-.90 $^{n.s.}$

Appendix 1. Scenarios used in studies 1-3

Studies 1 and 3

During a press conference last week, in response to a question about immigrant workers, a spokesperson of [Company Name] said:

Case 1 (Pro-immigrants): "As a company, we believe that all illegal immigrants cannot be asked to leave the country without considering how long they have already been living here. This is a matter of compassion. There should be some solution. This country has been benefited in many ways by the contributions of such illegal immigrants."

Case 2 (Contra-Immigrants): "As a company, we believe that all illegal immigrants have to be asked to leave our country irrespective of how long they have been living here. They violated the law of the land. There should be no amnesty. The rightful citizens of our country have been badly affected for far too long by such illegal immigration."

Case 3 (Control): "As a company, we believe our success is largely caused by our people. Everyone here contributes a part - their time, motivation, joy and work - to our processes and, in turn, to the creation of something great. Our activity makes sense for all of us and also for our customers. Therefore, everyone in this company can find meaning in their daily work."

Study 2

Case 1 (Pro-Choice, adopted from Mukherjee and Althuizen, 2020): According to reports, [Company Name] has been socially active and voiced their opinion regarding contemporary issues. For example, the company recently took a strong pro-choice stand in the reproductive rights debate. The company also financially contributes to the pro-abortion rights movement that actively supports Planned Parenthood.

Case 2 (Pro-Life, adopted from Mukherjee and Althuizen, 2020): According to reports, [Company Name] has been socially active and voiced their opinion regarding contemporary issues. For example, the company recently took a strong pro-life stand in the reproductive rights debate. The company also financially contributes to the anti-abortion rights movement that actively opposes Planned Parenthood.

Case 3 (No-Stance): According to reports, [Company Name] has been socially active but impartial on contemporary issues. For example, the company refrained from taking a stand in the reproductive rights debate. The company also decided to not financially contribute to the sociopolitical movements, such as the pro-life or pro-choice movements.

Appendix 2. Diagnosis of possible causes of endogeneity and techniques applied to counter them

Endogeneity cause	Hypothesised effects		
	Brand activism → Moral emotions*	Moral emotions → Brand attitude Moral emotions → Issue advocacy	Moderation by CBI
Omitted variable	▪ Participants were randomly assigned to an experimental condition** (Hill et al., 2021; Shadish, Cook, & Campbell, 2002), sample sizes were appropriate and manipulations were effective (Krause & Howard, 2003)	▪ Pre-brand attitude (studies 1-3), issue involvement (study 1) and age (study 3) were considered as control variables (Hill et al., 2021) ▪ In all studies, key omitted variables (gender, age and political ideology are evenly distributed across treatment groups (Hill et al., 2021)	
Simultaneity	▪ Participants were randomly assigned to an experimental condition** (Hill et al., 2021; Shadish et al., 2002), sample sizes were appropriate and manipulations were effective (Krause & Howard, 2003)	▪ It is considered proven that there is not a large gap (in time) between a stimulus and the emergence of moral emotions (Berkowitz & Harmon-Jones, 2004; Lerner, Goldberg, & Tetlock, 1998) ▪ Pre-brand attitude (studies 1-3), issue involvement (study 1) and age (study 3) were considered as control variables (Hill et al., 2021)	▪ Consumer-brand identification was measured before the participants were exposed to a scenario (and thus before they were expected to elicit moral emotions and rate post-brand attitude) ▪ Pre-brand attitude was considered as control variable (Hill et al., 2021) and measured before post-brand attitude
Measurement error	▪ Validated measurement models were used (Greco, O'Boyle, Cockburn, & Yuan, 2018), whose high validity was confirmed in the studies (see Table 2) (Hill et al., 2021) ▪ In all studies, the tested models passed Harman's single factor test (see Table 2) (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Podsakoff, MacKenzie, & Podsakoff, 2012; Siemsen, Roth, & Oliveira, 2010) ▪ The scenarios were either taken from an established study (study 2) or subjected to a pre-test		

Selection into sample	▪ mTurk covers a broad, but not the entire spectrum of the target population (Hill et al., 2021) ▪ Participation was limited to panellists with an HIT approval of 95% or higher (Goodman & Paolacci, 2017) ▪ Participants were randomly assigned to an experimental condition** (Hill et al., 2021; Shadish et al., 2002), sample sizes were appropriate and manipulations were effective (Krause & Howard, 2003) ▪ The share of those who clicked on the survey link but did not fill in the questionnaire or did not fill it in completely does not differ between the groups in any of the studies
Selection of treatment	▪ Participants were randomly assigned to an experimental condition** (Hill et al., 2021; Shadish et al., 2002), sample sizes were appropriate and manipulations were effective (Krause & Howard, 2003)

Paper 2

Who is more responsive to brand activism? The role of consumer–brand identification and political ideology in consumer responses to activist brand messages

Status: published at the *Journal of Product and Brand Management*

Co-Authors: Stefanie Wannow, Linda Marquardt, Jana S. Graubner and Alexander Haas

Own Share: 50%

Who is more responsive to brand activism? The role of consumer–brand identification and political ideology in consumer responses to activist brand messages

Abstract

Purpose – Through activism, brands participate in the socio-political controversies that shape our society today. Based on social identity theory, this study examines the moderating effects of consumer–brand identification (CBI) and political ideology in explaining consumer responses to brand activism. Furthermore, the role of perceived marginalization that can arise in case of consumer–brand disagreement is explored.

Design/methodology/approach – The hypothesized effects were tested in three experiments. Study 1 (n = 262) and study 2 (n = 322) used a moderation analysis, which was supplemented by a mixed design analysis with repeated measures in study 1. In study 3 (n = 383), the mediating effect of perceived marginalization by the brand was tested using a moderated mediation model.

Findings – The results show that strong CBI as well as a conservative ideology buffer the negative effects of consumer–brand disagreement on brand attitude and word-of-mouth intentions. In the case of agreement with a brand’s stance, no direct or interactive effects of brand activism on consumer responses occur. Perceived marginalization by a brand mediates the effects of brand activism.

Originality – This study extends the “love is blind” versus “love becomes hate” debate to the realm of brand activism and finds evidence for the former effect. It also contributes to the research on political consumption by highlighting the role of political ideology as an important boundary condition for brand activism. Perceived marginalization is identified as a relevant risk for activist brands.

Keywords: Brand activism, Brand attitude, Consumer–brand identification, Political ideology, Perceived marginalization

1. Introduction

Recently, an increasing number of brands have spoken up and even taken action that have contributed to divisive sociopolitical causes (Moorman, 2020). This phenomenon has been termed corporate or brand activism, which is “a purpose- and value-driven strategy” (Vredenburg *et al.*, 2020, p. 446), in which brands take a non-neutral and public stance on contested sociopolitical issues to achieve social change and economic goals (Hydock *et al.*, 2020). For instance, after the overturn of the constitutional right to an abortion by the Supreme Court, companies such as Disney, Airbnb, and Microsoft stated that they would cover travel expenses to help employees gain access to abortion procedures (Goldberg, 2022). Interestingly, companies headquartered in Democratic states were significantly more outspoken on supporting reproductive rights than companies headquartered in Republican states (Sonnenfeld *et al.*, 2022). This political divide is also visible in the population. Even though most of the US population is in favor of legal abortion rights at least under certain circumstances, significantly more Democrats classify themselves as pro-choice (88%) than pro-life (12%), while 70% of Republicans self-define as pro-life and 23% as pro-choice (Gallup, Inc., 2022). This example illustrates the politically charged society in which companies operate today.

With the rise of brand activism, researchers have started to investigate this emerging brand strategy and have focused mainly on its consequences for the brand, namely, on attitudinal and behavioral consumer responses (Klostermann *et al.*, 2021; Mukherjee and Althuizen, 2020) as well as on economic (i.e., market shares) and financial outcomes (Bhagwat *et al.*, 2020; Hydock *et al.*, 2020). In addition, there is initial evidence of the underlying psychological mechanisms of consumers’ responses, such as the mediating role of consumer–brand identification (CBI) and emotions (Garg and Saluja, 2022; Matos *et al.*, 2017; Mukherjee and Althuizen, 2020). Overall, the results of the existing studies indicate that companies should be highly cautious about brand activism due to an asymmetric effect that leads to stronger reactions by opponents (vs. supporters) of the brand’s stance (Bhagwat *et al.*, 2020; Mukherjee and Althuizen, 2020). Furthermore, the outcomes of brand activism are contingent on various factors. For instance, several studies have identified authenticity as a central factor for brands seeking to strengthen their brand equity and achieve social change (Ahmad *et al.*, 2022; Mirzaei *et al.*, 2022; Schmidt *et al.*, 2021). However, despite early laudable efforts to better understand boundary conditions of successful brand activism (e.g., Hydock *et al.*, 2020), research on potential moderating effects is still scarce.

To address these research gaps, this study investigates potential moderators (i.e., CBI and political ideology) and a novel mediator (i.e., perceived marginalization by the brand) from the perspective of social identity theory. Social identity theory is based on the premise that an individual's sense of self is derived from their group membership(s) (Tajfel and Turner, 1979). A group is defined as "a collection of individuals who perceive themselves to be members of the same social category" (Tajfel and Turner, 1979, p. 40). This social category may be determined by culture, sports, work, and any other formal or informal social entity to which one can develop a sense of identification and belonging (Bhattacharya and Sen, 2003; Lam *et al.*, 2010). Individuals define their identities based on their affiliation to these social entities (i.e., their ingroups) to strengthen and protect their self-identity (Tajfel and Turner, 1979).

In referring to a brand as a source of identity, the authors first seek to answer the question of how a pre-existing CBI shapes consumer responses to brand activism as suggested by Mukherjee and Althuizen (2020). Demonstrating belongingness to a valued brand is one of the multiple possibilities to build and express one's sense of self (Bhattacharya and Sen, 2003; Escalas and Bettman, 2005) and is commonly captured by CBI (Kuenzel and Halliday, 2008; Stokburger-Sauer *et al.*, 2012; Tuškej *et al.*, 2013). The current research ties into the long-running "love is blind" versus "love becomes hate" debate concerning the impact of CBI and related constructs on consumer reactions to adverse brand events (i.e., product failures and ethical transgressions) (e.g., Trump, 2014; Zhang *et al.*, 2020). Research findings are inconclusive as to whether existing consumer–brand relationships mitigate or exacerbate negative effects (Khamitov *et al*., 2020). On the one hand, attached consumers are motivated to justify the brand's actions to protect their cherished brand relationship. On the other hand, perceived misconduct could pose an excessive threat to their self-concept and therefore lead to a clear distancing from the transgressing brand. Given the plausible theoretical arguments for both effects and the lack of research in the context of brand activism, the authors propose and test two competing hypotheses (e.g., Chaker *et al*., 2019; Kaplan *et al*., 2007).

Second, this research aims to contribute to the question of how political ideology moderates the effects of consumer–brand (dis)agreement with a brand's stance. Identifying with a particular political ideology or party is another important source of one's self (Jung and Mittal, 2020; Weber *et al.*, 2023). Recent research has addressed the role of political ideology in the context of brand activism, but mainly focused on the divide between typical conservative versus liberal stances and the respective reactions of conservative and liberal consumers (e.g., Garg and Saluja, 2022). Although it seems logical that liberals (vs. conservatives) respond more (less) positively to typical liberal stances (e.g., strict gun control) and vice-versa, the strength of their responses to brand activism is not easy to predict. As an extension of the valuable study by Ketron *et al.* (2022), who were the first to find differences in responsiveness levels, this study examines whether there is any positive value of brand activism at all for liberals versus conservatives. While Ketron *et al.* (2022) compared the effects of disagreement versus agreement, this paper also explicitly integrated control groups or pre-measurements of dependent variables (i.e., brand attitude and word of mouth) in the studies to understand the effect of brand activism versus a non-activist situation on consumer responses. This comparison makes it possible to test the postulated asymmetric effects of brand activism, i.e., the presence of negative effects in the absence of positive effects, as suggested by Mukherjee *et al*. (2020), for different groups of consumers.

Third, the current study explores the role of perceived marginalization by the brand. Building on research on perceived societal marginalization (Bollwerk *et al.*, 2021), it is argued that in the current polarized environment, self-concept is shaped by moral and political beliefs (Jung and Mittal, 2020). Consequently, when consumers disagree with a particular stance taken by a brand, they feel devalued and excluded as part of a particular group that is marginalized by the brand. Thereby, this research aims to contribute to a better understanding of the underlying psychological mechanisms of consumer responses to brand activism while accounting for differences in perceived marginalization due to CBI and political ideology.

In the following sections, the research on the "love is blind" versus "love becomes hate" debate as well as on political ideology and consumer behavior is reviewed and placed in the context of brand activism. After introducing the proposed moderating effects, the role of perceived marginalization is explained. To test the hypotheses, three experimental studies are presented. The article concludes with a discussion of the main findings and theoretical and managerial implications and finally state limitations and future research directions.

2. Conceptual framework

2.1. CBI as a boundary condition of consumers' responses to brand activism

CBI

The theoretical foundation of CBI is the notion that consumers look for brands and products with which they can identify and express their self (Belk, 1988). The tendency of consumers to choose brands that match their current or desired identity has been established by multiple studies (e.g., Bhattacharya and Sen, 2003; Escalas and Bettman, 2005; Tuškej *et al.*, 2013). Through the means of suitable brands, consumers seek to construct and express their identity as well as demonstrate an affiliation to or demarcation from certain groups (Escalas and Bettman, 2005). Against this theoretical background, CBI has been defined as "a consumer's perceived state of oneness with a brand" (Stokburger-Sauer *et al.*, 2012, p. 407), namely, it is the degree of connectedness between one's own identity and the brand's identity (Davvetas and

Diamantopoulos, 2017). When a brand has the characteristics that a person identifies with, a sense of belonging to that brand is created (Lam *et al.*, 2010).

Brand activism typically contains relevant information about the brand's values, which is an important source of CBI (Tuškej *et al.*, 2013). According to Mukherjee and Althuizen (2020, p. 2), brand activism "provides consumers with an opportunity to assess the level of self-brand similarity in the context of moral judgments," thus enabling consumers to compare their own values to the brand's values. Recent studies have already established that consumer–brand disagreement on a socio-political issue harms CBI. In these studies, CBI was integrated as a response to brand activism (Garg and Saluja, 2022; Matos *et al.*, 2017; Mukherjee and Althuizen, 2020). Conceptualizing CBI as a moderator does not contradict this finding. On the one hand, marketing aims to influence attitudes and behavior; therefore, it makes sense to view these variables such as CBI as responses to a specific action. On the other hand, consumers' reactions are shaped by their existing attitudes and attachments, which are based on cumulative brand experiences of consumers (Hoyer *et al.*, 2018). Accordingly, past research has conceptualized CBI as both a moderating variable (e.g., Dalman *et al.*, 2019; Xiao and Lee, 2014) and as a mediating or dependent variable (e.g., He and Li, 2011; Stokburger-Sauer *et al.*, 2012). This research focuses on the moderating role of existing CBI based on Mukherjee and Althuizen's (2020, p. 785) suggestion to "study the influence of prior brand attachment on the effects of brand activism." Especially for well-known activist brands, it seems reasonable to assume that consumers have already formed a sense of identification—or a lack of it—that acts as a boundary condition. This understanding is also the basis of the "love is blind" versus "love becomes hate" research which will be outlined in the following section.

How existing consumer–brand ties affect responses to negative brand actions

How existing consumer–brand relationships affect individuals' reactions to certain (predominantly negative) marketing actions is controversially discussed in research (e.g., Khamitov *et al.*, 2020; Trump, 2014). When public figures or brands commit transgressions, they create a difficult situation for individuals who have strong emotional connections with them. People value their own moral standards and want to view themselves as morally upright, which means they avoid behaviors that might conflict with their moral beliefs. However, when people are attached to an individual or brand, they are also strongly inclined to protect their positive view of them. In a case where public figures or brands commit transgressions, a conflict between their motivation to support the transgressor and their need to maintain their moral values will be created (Bhattacharjee *et al.*, 2013; Lee *et al.*, 2020).

Thus, while scholars agree that consumer–brand relationships matter when consumers experience product or service failures or learn of ethical misconduct, the exact patterns of influence are not yet fully understood: "On the one hand, prior relationships can buffer negative impact and serve as a safety cushion; but, on the other hand, the existence of such relationships can amplify the negative impact and backfire" (Khamitov *et al.*, 2020, p. 530). Similarly, the literature review (see Appendix 1 for a full overview) illustrates the existence of both types of effects, with multiple studies showing the presence of one type of effect or the other. Additionally, some studies discovered both effects (Einwiller *et al.*, 2019; Ma, 2020; Wan *et al.*, 2011; Zhang *et al.*, 2020). The type of effect seems to be contingent on different moderating factors such as controllability and type of failure or transgression (Grégoire and Fisher, 2006; Saavedra Torres *et al.*, 2020; Trump, 2014), relationship type or norms (Aggarwal, 2004; Wan *et al.*, 2011), and the severity of the transgression (Einwiller *et al.*, 2006; Liu *et al.*, 2010; Schmalz and Orth, 2012).

In the specific context of brand activism, Mukherjee and Althuizen (2020) considered either buffer or amplifying effects to be possible when consumers hold views and values differing from those of the brand. On the one hand, they suggested that if there is a strong bond between consumers and brands, "the psychological cost of boycotting the brand will be high" (Mukherjee and Althuizen, 2020, p. 785). This could lead consumers to rationalize what they perceive as an immoral stance, thus mitigating potentially negative effects on brand attitudes and behaviors. On the other hand, the authors argued for the presence of a possible amplifier effect caused by higher expectations of strongly identified consumers. Thus, the question of whether strong CBI leads to one or the other effect is still largely open. Accordingly, two competing hypotheses are being developed and evaluated. Scholars have noted that examining competing hypotheses enhances objectivity (Armstrong *et al.*, 2001) and is particularly useful when prior knowledge allows different plausible justifications (Chaker *et al.*, 2019; Sawyer and Peter, 1983).

Competing hypothesis 1: CBI buffers the negative effects of brand activism

The buffering—or "love is blind"—effect suggests that consumers close to a brand are more tolerant in case of perceived brand transgressions. This effect is based on the notion that strongly connected consumers resist information that challenges their pre-existing brand attitudes and relationships (Ahluwalia *et al.*, 2000; Leak *et al.*, 2015). Connected consumers process information in a biased way and use different moral reasoning strategies to protect their positive view of a brand and avoid distancing themselves from it, which would come with high psychological costs (Mukherjee and Althuizen, 2020). These strategies include moral rationalization, namely, ignoring negative information, developing counterarguments, and justifying negative actions and moral decoupling, i.e., separating moral from performance judgments, which in turn prevent negative attitude change (Ahluwalia, 2002; Bhattacharjee *et al.*, 2013; Mukherjee and Altzuizen, 2020). This way, consumers protect themselves against identity threats (Cheng *et al.*, 2012; Leak *et al.*, 2015; White and Argo, 2009) and negative consumer responses (e.g., brand boycotts) are mitigated (Schmalz and Orth, 2012). In support of this argument, Dalman *et al.* (2019) proposed that strong emotional bonds make individuals more likely to seek external explanations in case of an unethical or immoral act done by someone close, leading to less harsh judgments. In their study, the authors found that brand identification is, overall, beneficial for a brand that has engaged in ethical misconduct.

In addition, Pullig *et al.* (2006) argued that prior brand attitudes incorporate manifold positive brand associations. When consumers are confronted with negative brand-related information, the existing brand knowledge serves as a basis for evaluation and may cause a halo effect (Leak *et al.*, 2015; Pullig *et al.*, 2006). This line of reasoning was also adopted by Leak *et al*. (2015). To the best of the author's knowledge, their research on consumer responses to managers' ideological stances represents the only study in the context of activism to examine the role of prior brand attitude. Suggesting a halo effect, the scholars found a 'love is blind' effect as positive pre-existing brand attitudes reduced consumers' negative reactions in case of disagreement with an ideological manager's stance. Using the topic of same-sex marriage, the researchers concluded that "the more a consumer likes a brand prior to exposure to a manager's position, the more likely negative shifts in brand attitude will be mitigated" (Leak *et al.*, 2015, p. 495). Based on this and related research evidence and the theoretical arguments provided, the following hypothesis can be derived:

> ***H1a***: The negative effect of consumer–brand disagreement on consumer responses (brand attitude and word-of-mouth (WoM)) is buffered by CBI.

Competing hypothesis 2: CBI amplifies the negative effects of brand activism

In contrast, other studies suggest that the opposite effect is to be expected (see Appendix 1). From a theoretical perspective, the amplifier—or "love becomes hate" effect—can be explained by the higher expectations consumers build up toward the brands they are committed to (Grégoire *et al.*, 2009; Ma, 2020; Mukherjee *et al.*, 2020; Trump, 2014). When a brand does not live up to these expectations, consumers feel let down. Similarly, brand transgressions can be experienced as a direct threat to a consumer's self-concept or even as a personal failure (Antonetti and Anesa, 2017; Cheng *et al.*, 2012; Kim *et al.*, 2020; Weitzl *et al.*, 2023). This self-threat and unfulfilled expectations can result in negative emotions such as disappointment, betrayal, shame, and anger, which motivate consumers to turn away from or even advocate against the company instead of justifying the company's moral misconduct (e.g., Grégoire and Fisher, 2008; Jain and Sharma, 2019; Weitzl *et al.*, 2023).

Scholars found that the "love becomes hate" effect is typically associated with severe fairness violations (Grégoire and Fisher, 2008), threats to communal (Aggarwal, 2004) or identifying consumer–brand relationships (Ma, 2020), and, most importantly, negative brand events that are highly self-relevant or represent ethical misconduct (Einwiller *et al.*, 2019; Saaveedra Torres *et al.*, 2020; Trump, 2014; Weitzl *et al.*, 2023). A possible explanation is the notion that moral transgressions are seen as particularly indicative of one's true character and, therefore, are difficult to ignore or justify (Einwiller *et al.*, 2006; Einwiller *et al.*, 2019; Folkes and Kamins, 1999; Lee *et al.*, 2020). Because the perception of one's own morality is a crucial aspect of the self-concept (Strohminger and Nichols, 2014), suspecting an immoral brand character is especially threatening to one's self. Therefore, it can be argued that highly attached consumers will show stronger reactions, which may go as far as a vehement rejection of a company that they feel no longer aligns with their moral principles (Bhattacharya and Sen, 2003; Einwiller *et al.*, 2019).

Consequently, in the context of brand activism, an amplification effect of CBI can also be argued because CBI typically deals with morally charged issues that involve ethical controversies. For example, when a brand speaks out against transgender rights, a consumer who supports these rights might consider the brand's stance as an ethical transgression. As outlined above, a strong CBI means that the brand's activities will be perceived as self-relevant and self-defining and, consequently in case of consumer–brand disagreement, as self-threatening because the consumer's own identity is suddenly being questioned (Einwiller *et al.*, 2019; Ma, 2020; Trump, 2014). For consumers who disapprove of an activist position but do not identify with a brand, this personal relevance is not felt, thus the brand's actions have a scant influence on the self and subsequent brand-related responses. Therefore, it can be argued that CBI will lead to a stronger negative effect of consumer–brand disagreement on brand attitude as consumers seek to distance themselves from potentially stigmatizing brand actions.

***H1b*:** The negative effect of consumer–brand disagreement on consumer responses (brand attitude and word-of-mouth (WoM)) is amplified by CBI.

2.2. The role of political ideology and perceived marginalization by the brand

Political ideology and consumer behavior

The second main objective of this paper is to better understand the differences between conservative versus liberal consumers when responding to brand activism. The question of how politics and individual political ideology shape consumer behavior and how companies should respond to politicized consumers has attracted considerable interest in recent years (e.g., Fernandes, 2020; Jung and Mittal, 2020; Pecot *et al.*, 2021). Politics today plays a great role in various domains of our lives, which includes the domain of consumption (Korschun *et al.*, 2020). Political debates on diverse issues, such as LGBTQ+ rights, abortion, or gun control, have increased in number and fierceness, reflecting the polarized state of many societies. Despite varying viewpoints about the exact state of the current polarization, most researchers agree on a substantial divide between partisan groups (Weber *et al.*, 2023), especially in the United States. Americans have recently not only become increasingly divided in terms of their political beliefs and values, they have also developed a growing aversion toward the opposite side (Iyengar *et al.*, 2019).

In times of polarization, political ideology has become a salient part of consumer identity (Jung and Mittal, 2020; Weber *et al.*, 2023). In other words, association with political groups and partisanship are another important source of consumers' social identity today (Mason, 2015). Due to its salience, "one can expect consumers to behave in ways that reinforce their political identity" (Jung and Mittal, 2020, p. 62). Research has already established the impact of consumers' political identities on their behavior, such as boycotting and other forms of "anti-consumption" (Fernandes, 2020; Northey *et al.*, 2020; Pecot *et al.*, 2021). This should apply in particular to brand activism, which deals with highly ideologically driven topics that polarize opposing partisan groups as already indicated by initial research (Garg and Saluja, 2022; Ketron *et al.*, 2022; Weber *et al.*, 2023). In addition, research in related areas has demonstrated the relevance of political identity in the context of controversial or negatively perceived brand actions (see Appendix 2 for an overview).

To capture political identity, this study focuses on the political ideology of individuals, which is commonly conceptualized as an individual's position on the political spectrum ranging from conservatism to liberalism (Hambrick and Wowak, 2021; Jost *et al.*, 2009; Weber *et al.*, 2023). This self-categorization on the liberal-to-conservative spectrum or, in other countries, the left-to-the-right continuum (Hambrick and Wowak, 2021), has been found to be a reliable predictor of political attitudes (Day *et al.*, 2014; Mason, 2015). Political ideology can be defined as a "set of core values and beliefs about how society should be organized" (Pecot *et al.*, 2021, p. 62). By identifying with a certain party that stands for specific values and ideals, individuals express their political ideology (Antonetti and Anesa, 2017). The political values of liberals (as opposed to those of conservatives) are, for example, characterized by their preference for social change over tradition, their rejection rather than acceptance of inequality, and their emphasis on fairness and harm avoidance over authority, loyalty, and purity (Day *et al.*, 2014; Jost *et al.*, 2009; Pecot *et al.*, 2021; Fernandes, 2020).

These value priorities influence how consumers respond to political statements of brands: given that brand activism deals with polarizing socio-political issues "that induce both support and opposition, it follows that an individual's consumer attitude toward the brand will depend on their political beliefs" (Klostermann *et al.*, 2021, p. 781). Naturally, activism that is representative of a political ideology will receive more support from the respective partisans, while consumers punish brands they oppose politically (Duman and Ozgen, 2018). For example, a pro-life statement is expected to receive backlash mainly from liberals (Garg and Saluja, 2022; Weber *et al*., 2023). Beyond this support by political camps for "their" positions on socio-political issues, this research focuses on the potentially different responsiveness of liberals versus conservatives to activist messages.

Political ideology as a moderator

Scholars have suggested that liberal consumers are likely to include political considerations in their consumption decision (e.g., Fernandes, 2020; Pecot *et al*., 2021). In their recent review, Copeland and Boulianne (2022) asserted that consumers on the left or liberal side of the political spectrum are more inclined to engage in political consumerism compared to conservative consumers, which is the "deliberate purchase or avoidance of products, goods, or services for political reasons" (Copeland and Boulianne, 2022, p. 3). Therefore, liberal consumers seem to be more sensitive to political messaging by brands.

One general reason for that sensitivity could be the cognitively more rigid and less deliberate decision-making style of conservatives (Angle *et al.*, 2017; Jung and Mittal, 2020) that might lead to more stable brand evaluations and behaviors. Liberals are characterized as being more susceptible to the influence of novel information and therefore more malleable in their views (Angle *et al.*, 2017). In other words, a single event may not alter the evolution of attitude toward the brand among conservatives as much as it will among liberals. Further, there is initial evidence that conservatives are more focused on utilitarian brand benefits whereas liberals care more about abstract image attributes (i.e., image and brand personality dimensions) (Jung and Mittal, 2020). As activism is not linked to the utilitarian aspects (such as quality and performance) of a brand but is instead linked to abstract values, it may be less relevant for the attitude formation of conservatives.

The fact that liberal (vs. conservative) activism is more common (Hambrick and Wowak, 2021) could further explain why conservative consumers might not react as strongly to political brand messages. For example, in Klostermann *et al.*'s (2021) overview of 106 corporate political advocacy events, liberal events formed the great majority. Therefore, conservatives are used to their political beliefs not being reflected in activist brand communication. Similarly, they exhibit system justification tendencies (Jost *et al.*, 2008; Napier and Jost, 2008), which include accepting and rationalizing existing institutions and having favorable attitudes toward the overarching social and political structure despite the fact that it may be detrimental to them (Jost *et al.*, 2009, Jung *et al.*, 2017). These characteristics may serve as a barrier against liberal-leaning brand activism, which conservatives may perceive as a normal business practice (Bailey and Phillips, 2020; Napier and Jost, 2008).

A first scientific study supports the different response patterns of liberal versus conservative consumers. In their research, Ketron *et al.* (2022) identified conservative consumers as less sensitive to brand activism. These scholars explained the strong reactions from liberal consumers to be due to the great emphasis the liberal consumers place on corporate activism and, therefore, the higher weight they attribute to political stances in their brand attitude formation. A study of managers and MBA students revealed comparable results (Bailey and Phillips, 2020). While Democrats punished companies that contradicted their personal values, Republicans' brand attitudes were not affected by a liberal brand's stance. In summary, initial evidence suggests that liberal consumers exhibit stronger responses than conservative consumers toward brand activism. Therefore it is hypothesized:

H2**:** The negative effect of consumer–brand disagreement on consumer responses (brand attitude and word-of-mouth (WoM)) is amplified by a liberal political ideology.

2.3. Asymmetric effects of brand activism

Building on the established negativity bias, i.e., "the greater power of bad events over good ones" (Baumeister *et al.*, 2001, p. 323), a disproportionally negative effect of brand activism is expected. For several decades, psychologists have found evidence that consumers generally react more strongly to negative than to positive information and events (Baumeister *et al*., 2001; Liu *et al*., 2010; Sen and Bhattacharya, 2001; Taylor, 1991). This effect has been termed negativity bias or positive–negative asymmetry effect and is assumed to be caused by factors such as high levels of perceived relevance, salience, and diagnosticity of negatively valenced information (Baumeister *et al.*, 2001). Research on the adjacent topic of corporate social responsibility (CSR) supports this negativity bias. For example, Sen and Bhattacharya (2001) found that only consumers who strongly support a particular issue react favorably to positive CSR information, whereas all consumers react unfavorably to negative CSR information. Similarly, consumers are more likely to engage in boycotting due to unethical brand actions than to engage in "buycotting" (Trudel and Cotte, 2009).

For brand activism, consumer reactions have also been found to be asymmetric, thus supporting the notion of the negativity bias (Hydock *et al*., 2020; Jungblut and Johnen, 2021; Mukherjee and Althuizen, 2020). For instance, Mukherjee and Althuizen (2020) identified the "asymmetric effect of brand activism": while agreement with an activist message had no effect on consumers' responses, disagreement had a detrimental effect on brand attitudes, intentions, and behaviors. According to these scholars, this effect is based on different moral judgments elicited by brand activism. In the case of agreement, the brand's moral actions seem to be taken for granted and therefore do not lead to an improvement in brand attitudes and behaviors. Given that brand activism includes moral issues—and doing what is ethical 'just' is seen as a critical component of one's self (Einwiller *et al.*, 2019)—customers already expect the brand to share their point of view, as this is what they believe to be "correct" moral behavior (Mukherjee and Althuizen, 2020). In turn, people are unlikely to reward a brand for expressing a stance that meets their default moral expectations. Therefore, it is assumed that consumers do not reward brands for brand activism regardless of their level of CBI and their political ideology:

H3a: Consumer–brand agreement has no effect on consumers' brand-related responses (brand attitude and WoM) independent of their level of CBI.

H3b: Consumer–brand agreement has no effect on consumers' brand-related responses (brand attitude and WoM) independent of their political ideology.

2.4. The role of perceived marginalization as a mediator between brand activism and consumers' brand-related responses

Finally, it is proposed that consumers' identities could be threatened by perceptions that a brand does not consider them and their values as relevant, thus leading to a sense of marginalization. Marginalization, as a general concept, can be defined as the "peripheralization of individuals and groups from a dominant, central majority" (Hall, 1999, p. 89). Marginalized groups are in some way intentionally excluded and rejected from the broader society (Betts and Hinsz, 2013). Minorities, for example, those based on religion or race, often suffer from marginalization and underrepresentation, and, thus, from a lack of power, control, and resources (Gardberg and Newburry, 2013). In consumer research, the concept of marginalization has so far attracted limited research interest. Existing studies have explored the consumer behavior of marginalized groups, for example, transgender (Crosby *et al.*, 2015) or Islamic consumers (El-Bassiouny, 2014). Other research has examined the perception of stereotypical advertising models by marginalized groups (Rasmussen *et al.*, 2021).

In the field of brand activism, scholars have already proposed that consumers can become alienated by controversial brand actions (Mirzaei *et al.*, 2022; Vredenburg *et al.*, 2020), but without exploring that phenomenon in more detail. Alienated consumers typically feel separated from existing norms and values and lack identification with the marketplace and its actors (Mady, 2011). In the given context, consumers might be afraid that their needs and values will not be adequately addressed or that they will be excluded from the brand if they disagree with its position (Key *et al.*, 2021). Because marginalization reflects intentional social exclusion and rejection, it seems reasonable to apply this concept to brand activism and assess its impact on consumer–brand relationships.

While marginalization can be defined using objective marginalization criteria (e.g., access to education and health offers and socioeconomic status) (Bollwerk *et al.*, 2021), the current study focuses on the concept of perceived marginalization, which is rooted in the already introduced social identity theory (e.g., Betts and Hinsz, 2013; Jetten *et al.*, 2001). As outlined, social identities may be threatened by perceived—and not necessarily objectively given—inadequate group recognition, social rejection, and injustice (Bollwerk *et al.*, 2021; Papaoikonomou *et al.*, 2016; Rasmussen *et al.*, 2021). If a social group is collectively devalued, the group members can develop feelings of exclusion (Bett and Hinsz, 2013). To capture this experience, the construct of perceived societal marginalization was developed recently, and it reflects the "subjective and cognitive perception that one's own social group is unappreciated and treated as unimportant by society" (Bollwerk *et al.*, 2021, p. 4). Thus, the perception that the society or the marketplace devalues one's in-group leads to a feeling of discrimination (Jetten *et al.*, 2001; Papaoikonomou *et al.*, 2016).

The concept of marginalization is transferred to feelings of exclusion caused by activist brands because there is no brand-related construct in consumer research yet. Because engaging in activism is an intentional decision by a brand, the criterion of intentionality as a constituting characteristic of (perceived) marginalization is given (Betts and Hinsz, 2013). Moreover, activism always means taking sides in controversy and, accordingly, making an explicit statement against an opposing position. Those who disagree with a brand's stance may consequently feel discriminated against and excluded by the brand. Therefore, brand activism would be a source of devaluation for an individual and his or her in-group, which the brand has placed in an "outsider" position. In other words, this individual will feel marginalized. In turn, negative feelings of marginalization are likely to harm consumers' brand-related responses leading to the following hypotheses:

> ***H4*:** Perceived marginalization by the brand mediates the effect of consumer–brand disagreement on a consumer's brand-related response (i.e., brand attitude).

Figure 1 illustrates the derived conceptual model.

[Insert Figure 1 about here]

3. Empirical studies

3.1. Pilot study and an overview of main studies

A pilot study was created to find two divisive socio-political issues and two existing, well-known brands as message-senders for the scenarios in the main studies. Therefore, a desk analysis was used evaluating secondary sources and statistics to find polarizing issues (Moorman, 2020; Mukherjee and Althuizen, 2020; Vredenburg *et al.*, 2020) and brands that could realistically engage in brand activism. In the second step, two qualitative expert interviews were conducted to reduce the number of options related to both the issue and the brand. Following the experts' recommendations, the topics of illegal immigration and abortion were chosen and the choice of viable brands was reduced to 10. In the third step, to the target was to select two brands for the experiments. Accordingly, a survey with 80 respondents from Amazon mTurk was conducted. Participants were randomly assigned to one of four scenarios (pro-immigrant, anti-immigrant, pro-life, and pro-choice). They rated how realistic they perceived each of the ten brands to engage in brand activism and express the respective opinion.

For five brands, a comparison of alternative positions on an issue showed that the positions were not perceived as being equally realistic. Out of the remaining brands, a well-known hypermarket chain and a DIY retailer were chosen. Both brands were rated with the highest realism values regarding both possible stances on the respective issue (e.g., pro-life and pro-choice).

Three experimental studies were conducted to test the proposed relationships. Study 1 used a moderation analysis to analyze the moderating effects of CBI and political ideology on brand attitude change. Furthermore, mixed-design analyses were used to measure the proposed asymmetric effect of brand activism. Based on a between-subjects design and a moderation analysis, study 2 served to replicate the findings of study 1 using WoM as another relevant dependent variable. In study 3, the hypothesized mediation effect of perceived marginalization was tested. To ensure the robustness of mediation paths and explore the role of CBI and political ideology in more detail, an additional moderated mediation model was employed.

3.2. Study 1

Participants and procedure

Study 1 aimed to examine the moderating effects of CBI and political ideology on the relationship between consumer–brand disagreement and consumers' brand-related responses (i.e., brand attitude). Therefore, the ratings of pre- and post-treatment brand attitudes were compared. Based on the results of a pre-test, the issue of illegal immigrants was chosen and a well-known U.S. hypermarket chain was selected as the sender of the brand activist statement (see Appendix 3 for scenarios). In total, 262 adults from Amazon mTurk (45.0% female; $M_{age} = 39.06$, $SD_{age} = 10.88$) participated in the study.

Measures

Before being exposed to one of the scenarios, respondents had to rate their pre-existing brand attitude and pre-existing CBI related to the chosen hypermarket brand. Next, attitude towards the issue (i.e., illegal immigrants) was measured by the level of agreement with the statement "all illegal immigrants should be asked to leave the USA irrespective of how long they have been here" (1 = "fully disagree" and 7 = "fully agree") (Mukherjee and Althuizen, 2020). Following this assessment, participants were randomly assigned to either a "pro-" or "anti-immigrant" scenario that was adopted from Mukherjee and Althuizen's (2020) study that included a public statement by a brand either in favor of or against mass deportation of unauthorized immigrants. Corresponding to the respective scenario, respondents' pre-activism attitude towards unauthorized immigrants was coded from 1 = "fully agree with the company's stance" to 7 = "fully disagree with the company's stance." For example, respondents who reported their full approval of mass deportations and were exposed to a brand's statement against mass deportations received a score of 7 ("fully disagree") on the newly created disagreement scale. CBI was measured using the scale developed by Stokburger-Sauer *et al.* (2012) with a seven-point Likert scale (1 = "fully disagree" and 7 = "fully agree"). Brand attitude was captured through a seven-point semantic differential by Mukherjee and Althuizen (2020). Moreover, political ideology was assessed through a semantic differential ranging from 1 = "very liberal" to 7 = "very conservative" (Davidson and Theriault, 2021). In the sample, 32.4% of the respondents considered themselves as (rather) liberal and 58.0% as (rather) conservative, while the rest identified as politically neutral (M = 4.39; SD = 2.01). All values regarding internal consistency as well as discriminant and convergent validity were above the recommended thresholds (Hulland *et al.*, 2018) (see Table 1).

[Insert Table 1 about here]

Results

To test the moderating impact of CBI and political ideology on the effect of consumer–brand disagreement on brand attitude change (i.e., the difference between the pre- and post-measurement of brand attitude), a moderation analysis (PROCESS model 2; Hayes, 2021) was used including pre-activism brand attitude as a covariate. Results revealed that a higher level of disagreement led to a more negative post-activism brand attitude (b = -0.61; $p < 0.001$, 95% CI [-0.85; -0.38]). The interaction between disagreement and CBI was significant (b = 0.05, $p < 0.05$, 95% CI [0.006; 0.10]). Likewise, the interaction between disagreement and political ideology was significant (b = 0.04, $p < 0.05$, 95% CI [0.001; 0.08]). The interaction terms indicate a positive moderating effect of both CBI and a conservative ideology buffering the negative effect of disagreement on brand attitude change.

Subsequently, spotlight analyses (Spiller *et al*., 2013) were used to compare weak and strong CBI (-1 SD: 3.49 and +1 SD: 6.43) for liberals and for conservatives (-1 SD: 2.38 and +1 SD: 6.40). In the case of a weak CBI, a stronger disagreement led to a significant reduction of brand attitude for liberals (b: -0.33, $p < 0.001$, 95% CI [-0.44; -0.23]) as well as for conservatives (b: -0.18, $p < 0.05$, 95% CI [-0.33; -0.03]). In the case of a high CBI, the effect was still significantly negative for liberals (b: -0.18, $p < 0.05$, 95% CI [-0.32; -0.04]) but was not significant for conservatives ($p = 0.70$).

In sum, the smaller decrease in brand attitude among strongly (vs. weakly) identified consumers reflects the buffering or “love is blind” effect and, thus, supports *H1a*. Furthermore, the negative effect was stronger for liberals than for conservatives, thereby supporting *H2*.

To examine the hypothesized asymmetric effect of brand activism regardless of the level of CBI (*H3a*), a mixed ANOVA with CBI (strong vs. weak) and consumer–brand agreement (agree vs. disagree) as between-subjects factor and brand attitude as a repeated measure was used, controlling for political ideology. To create a categorical independent variable that distinguishes between respondents who agreed with the brand’s stance versus those who disagreed, the authors followed the procedure used by Mukherjee and Althuizen (2020):

Participants indicated their approval of the statement "all illegal immigrants should be asked to leave the USA irrespective of how long they have been here" (1 = "fully disagree"; 7 = "fully agree") (Mukherjee and Althuizen, 2020). For the anti-immigrant scenario, respondents who chose a value of 3 or less were defined as disagreeing, whereas those who chose a rate of 5 or more were defined as agreeing. For the pro-immigrant scenario, this system was inverted. Second, to distinguish between weakly versus strongly identified respondents, a median split had to be applied. Detailed results of the mixed ANOVA are shown in Table 2.

[Insert Table 2 about here]

Results reveal that attitude change depended on consumers' (dis)agreement because the interaction between disagreement and brand attitude was significant ($F(1, 222) = 18.17$, $p < 0.001$). Furthermore, the interaction between disagreement, CBI, and brand attitude was significant ($F(1, 222) = 4.20$, $p < 0.05$). Thus, the effect of (dis)agreement on brand attitude change depended on the level of CBI.

Estimated marginal means (see Figure 2) showed that brand attitude did not change (from pre- to post-brand activism) in the case of agreement ($p = 0.37$), but was significantly reduced in the case of disagreement ($F(1, 222) = 30.50$, $p < 0.001$). By including the differentiation of weak versus strong CBI, estimated marginal means showed that agreement with a brand's stance led to a marginally significant increase in brand attitude for consumers with weak CBI ($F(1, 222) = 2.99$, $p < 0.1$) but it was not significantly different for consumers with strong CBI ($p = 0.57$). In sum, this finding supports *H3a* and the asymmetric effect of brand activism. Yet, the marginally positive result for consumers with weak CBI indicates that there might be a potential for brand activism to increase brand attitudes to a limited extent (against a disproportionally large negative potential).

[Insert Figure 2 about here]

Second, to examine *H3b*, CBI was exchanged with political ideology, which distinguished between liberals (i.e., values between 1 and 3 on the 7-point Likert scale) and conservatives (i.e., values between 5 and 7 on the 7-point Likert scale). Correspondingly, CBI was used as control variable. Again, the interaction between disagreement and brand attitude was significant ($F(1, 201) = 23.43$, $p < 0.001$), and the three-way interaction between disagreement and political ideology and brand attitude was significant ($F(1, 201) = 10.37$, $p < 0.005$).

Estimated marginal means showed that brand attitude did not change (from pre- to post-brand activism) in the case of agreement ($p = 0.25$) but was significantly reduced in the case of disagreement ($F(1, 201) = 40.22$, $p < 0.001$). Moreover, estimated marginal means of the differentiation of liberals versus conservatives showed that agreement with a brand's stance did not lead to a brand attitude change either for liberals ($p = 0.16$) or for conservatives ($p = 0.98$), whereas disagreement led to a significant decrease of brand attitude for both liberals ($F(1, 201) = 39.45$, $p < 0.001$) and conservatives ($F(1, 201) = 4.64$, $p < 0.05$).

In sum, these results support the asymmetric effect of brand activism (*H3a and b*) as asserted by previous studies (Mukherjee and Althuizen, 2020; Hydock *et al.*, 2020): no positive effect of brand activism could be found in the case of consumer–brand agreement, irrespective of pre-existing CBI or political ideology.

3.3. Study 2

Participants and procedure

Study 2 aimed to examine a behavioral consumer response, i.e., WoM. Moreover, a neutral, non-activist statement was included as a control group. Therefore, another sample of 354 respondents from Amazon mTurk (35.9% female; $M_{age} = 36.87$, $SD_{age} = 12.43$) completed an online questionnaire in exchange for financial compensation. Participants were randomly assigned to one of three experimental conditions. As in study 1, the brand either took a "pro-immigrant" (i.e., against mass deportations) or an "anti-immigrant" (i.e., in support of mass deportations) position. As a third scenario, a non-activist PR statement about the company's market position and success served as control group (see Appendix 3). The classification of agreement and disagreement from Mukherjee and Althuizen (2020) was used again, as in study 1. In total, 105 respondents agreed with the brand's statement and 108 disagreed. The control group consisted of another 109 participants. Thirty-two participants who gave a rating of 4 (neutral) were excluded from the analyses because the sample size was considered too small for a moderation analysis.

Measures

Consumer–brand disagreement, CBI, and political ideology were measured using the same scales as in study 1. As outcome, WoM was assessed using three items from Xie *et al.*, 2019. Of the final sample of 322 respondents, 31.1% described themselves as (rather) liberal and 58.7% as (rather) conservative, while the rest said they were politically neutral (M = 4.53; SD = 2.03). All included constructs exhibited sufficient reliability and validity (see Table 1).

Results

To measure the effects of the brand activism condition (i.e., agreement, disagreement, and control) and the two moderators CBI and political ideology on WoM, a moderation analysis (PROCESS model 2, Hayes (2018)) was used including pre-brand attitude as a covariate. Consumer–brand agreement served as the baseline condition. Results revealed significant effects of disagreement (vs. agreement) ($b = -3.00$, $p < 0.001$, 95% CI [-3.93; -2.07]) and CBI ($b = 0.22$, $p < 0.005$, 95% CI [0.08; 0.36]) on WoM, but political ideology had no direct impact on WoM ($p = 0.56$). Furthermore, the interaction between disagreement (vs. agreement) and

CBI on WoM was significant ($b = 0.36$, $p < 0.001$, 95% CI [0.18; 0.54]). Likewise, the interaction between disagreement (vs. agreement) and political ideology on WoM was significant ($b = 0.15$, $p < 0.05$, 95% CI [0.02; 0.29]).

Spotlight analyses (Spiller *et al.*, 2013; see Fig. 3) were used to examine the effects at weak versus strong CBI (i.e., weak = -1 SD: 3.35 and strong = +1 SD: 6.49) for different political ideologies (i.e., liberals = -1 SD: 2.50 and conservatives = +1 SD: 6.56). For liberals with weak CBI, disagreement led to a significantly lower WoM than agreement ($b = -1.41$, $p < 0.001$, 95% CI [-1.82; -1.00]), as it did for conservatives with weak CBI ($b = -0.78$, $p < 0.005$, 95% CI [-1.31; -0.25]). In contrast, for liberals with strong CBI, the effect was not significant ($p = 0.29$), but it was marginally significant for conservatives with strong CBI ($b = 0.35$, $p < 0.1$, 95% CI [-0.05; 0.75]).

In sum, CBI was found to buffer the negative effects of consumer–brand disagreement on brand attitude. Therefore, evidence was found again in support of the buffer effect (*H1a*) and against the amplifier effect (*H1b*). Moreover, the results confirm the moderating role of political ideology in the context of brand activism. In support of *H2*, liberals, compared to conservatives, showed stronger negative reactions to brand activism they disagreed with.

Finally, to examine the proposed asymmetric effect of brand activism (*H3a and b*), effects of the control group versus the agreement group were assessed. As expected, the control group did not exhibit different WoM intentions than the agreement group ($p = 0.97$), and the interaction effects of the control group and CBI ($p = 0.41$) and of the control group and political ideology ($p = 0.72$), respectively, on WoM were not significant. The implication is that brand activism had no effect on WoM in the case of consumer–brand agreement (vs. control), independent of consumers' CBI or political ideology. Thus, the asymmetric effect of brand activism *(H3a and b)* could be supported.

[Insert Figure 3 about here]

3.4. Study 3

Participants and procedure

Study 3 examined the role of people's perception of being marginalized by the brand's stance as an underlying factor determining brand-related responses. In this study, the topic of abortion rights and a large and well-known U.S.-based DIY retailer brand were selected.

In total, 430 respondents from Amazon mTurk (34.0% female; M_{age} = 36.80; SD_{age} = 10.24) took part in this study. As activism scenarios, either a "pro-life" (against abortion rights) or "pro-choice" (supporting abortion rights) position was used. Moreover, a third scenario was included, where the brand explicitly denied taking an activist position (i.e., "no stance") (see Appendix 3).

Measures

Respondents reported their consumer–brand disagreement, CBI, and brand attitude on the same scales used in previous studies. Furthermore, to measure consumers' perceived marginalization by the brand, Bollwerk *et al.'s* (2021) societal marginalization scale was used. This scale is comprised of a political, cultural, and economic dimension. In this research, the focus was on perceived political marginalization because of the sociopolitical dimension of the stance taken by the brand in the study. Further, the items were adapted to fit the brand context. All included constructs exhibited sufficient reliability and validity (see Table 1). In this study, 47 respondents were excluded because they declared a neutral position toward the issue and could thus not be assigned to the "agree" or "disagree" group. The remaining sample of 383 respondents was diverse in terms of political ideology: 37.6% described themselves as (rather) liberal and 52.5% as (rather) conservative while the rest were politically neutral (M = 4.25; SD = 2.19).

Results

To test the mediation hypotheses (*H4*), a simple mediation model (PROCESS Model 4; Hayes, 2021) was first used with the brand activism condition (agreement, disagreement, no stance) as the independent variable, perceived marginalization by the brand as the mediator, and brand attitude as the dependent variable. Pre-brand attitude and political orientation were included as covariates and the agreement group served as the baseline comparison again. First, consumer–brand disagreement had a significant effect on perceived marginalization ($b = 0.88$, $p < 0.001$, 95% CI [0.50; 1.26]), which in turn decreased brand attitude ($b = -0.13$, $p < 0.001$, 95% CI [-0.17; -0.03]). In line with the hypothesized mediation effect, disagreement exerted a significant indirect effect via perceived marginalization on brand attitude ($b = -0.09$, 95% CI [-0.20; -0.06]). In addition, the direct effect of consumer–brand disagreement on brand attitude was significant ($b = -0.67$, $p < 0.001$, 95% CI [-0.93; -0.41]), indicating a partial mediation. Overall, these results support *H4*.

Moreover, the “no stance” scenario (vs. agreement) exerted an overall negative influence on brand attitude also, which was caused by a significant direct effect ($b = -0.42$, $p < 0.005$, 95% CI [-0.67; -0.17]). However, the indirect effect of “no stance” on brand attitude via perceived marginalization was not significant ($b = -0.04$, 95% CI [-0.12; 0.009]).

Second, to account for the identified moderation effects and test the robustness of the main mediation effects, a moderated mediation model (PROCESS model 10; Hayes, 2021) was conducted with CBI and political ideology as moderators and pre-brand attitude as a covariate (see Fig. 4). The effects of consumer–brand disagreement (vs. agreement) on perceived marginalization ($b = 2.53$, $p < 0.001$, 95% CI [1.51; 3.54]) and from the latter on brand attitude ($b = -0.13$, $p < 0.001$, 95% CI [-0.19; -0.06]) remained significant, as did the direct effect ($b = -1.58$, $p < 0.001$, 95% CI [-2.29; -0.88]). Interestingly, the interaction between disagreement and CBI had no significant effect on perceived marginalization ($p = 0.21$), but a significant effect on brand attitude ($b = 0.21$, $p < 0.005$, 95% CI [0.07; 0.35]). In contrast, the interaction between disagreement and political ideology had a significant effect on perceived marginalization (*b* = -0.25, *p* < 0.01, 95% CI [-0.44; -0.06]), but no significant effect on brand attitude (*p* = 0.97). Appendix 4 contains further details regarding direct and indirect effects at different levels of the moderators. Thus, CBI buffers the direct negative effect of disagreement on brand attitude, while a conservative ideology reduces its negative impact on brand attitude through perceived marginalization.

In sum, the results support findings of the previous studies (and *H2*): the negative effect of consumer–brand disagreement is stronger for liberal consumers than for conservatives even though conservatives perceive a higher level of marginalization on average (irrespective of whether they agree or disagree with the brand).

[Insert Figure 4 about here]

Regarding the deliberate no-stance message, results showed no major differences between the agreement and the no-stance group. The no-stance group (vs. agreement) exerted no significant effect on perceived marginalization (*p* = 0.49) or on brand attitude directly (*p* = 0.55). However, the interaction of no-stance and CBI on marginalization was significant (*b* = 0.27, *p* < 0.05, 95% CI [0.06; 0.47]). Likewise, the interaction of no-stance and political ideology on marginalization was significant (*b* = -0.30, *p* < 0.001, 95% CI [-0.47; -0.13]). Thus, CBI and a liberal ideology (both) increase feelings of marginalization on exposure to a deliberate "no stance" message.

4. Discussion

This study aimed to examine how two important dimensions of social identity, namely, pre-existing CBI and political ideology, shape consumer responsiveness to brand activism. A further goal was to explore the mediating role of perceived marginalization by the brand as an underlying factor of consumer responses to brand activism. Based on three experiments, the authors found that both strong CBI and a conservative ideology buffer against the negative effects of consumer–brand disagreement on brand-related responses. However, consumers' approval of the sociopolitical stance did not have any effects independent of the level of identification or political ideology. Moreover, consumers feel marginalized by activist brands when they disagree with the brands' stance, and this perception is detrimental to their attitude toward the brand.

Theoretical contributions

This research offers three major contributions to theory. First, it adds to the literature on negative brand events and the long-running "love is blind" versus "love becomes hate" debate. Results show that CBI can act as a buffer even when an issue is morally charged. Researchers have argued that threats to the moral self, which are particularly pronounced for connected consumers, make consumers turn their back on the brand (Einwiller *et al.*, 2019; Trump, 2014). However, this pattern did not emerge in this research; instead, the opposite was observed—strong CBI prevents negative consumer reactions from occurring. The buffering effect was detected in all three studies: it appears that the psychological costs for terminating a strong consumer–brand relationship are inordinately high vis-à-vis the benefit of upholding one's moral values by condemning and "punishing" a brand for the perceived transgression.

Therefore, it seems that different moral issues evoke different response patterns and self-protection mechanisms by consumers and the "love becomes hate" effect cannot be generalized to all moral issues. It is possible that the topics chosen and the nature of the brand's engagement (i.e., a single brand statement and lack of coordinated actions) were not perceived as such severe moral violations that as to turn consumers away from the brand. For instance, Ahmad *et al.* (2022) found that for high-equity brands, financial (vs. non-financial or rhetorical) commitment leads to stronger perceptions of authenticity and brand love. An emphasis on financial support for a given topic and consequential actions are likely to make a strong impression on consumers and signal sincere engagement (Bhagwat *et al*., 2020), which could also lead to a more extreme response from attached consumers who oppose the brand's stance. When a brand continuously invests in promoting a controversial stance, it might become impossible to justify the divide between brand and consumer, and this could hit strongly identified consumers particularly hard. In other words, full commitment to an activist cause would leave no room for moral rationalizing or decoupling but will rather lead to moral coupling with all its detrimental consequences for the brand (e.g., Mukherjee and Althuizen, 2020). On the contrary, when consumers perceive an isolated activist statement just as a marketing message without proof of sincere commitment to the issue, it will be easier for the identified consumers to condone it. Further, the exact source of the activist statement allows consumers to draw conclusions about the sincerity and possibly the severity of the brand's action (Bhagwat et al., 2020; Mukherjee and Althuizen, 2020).

An alternative explanation for the buffer effect could be rooted in consumers' awareness of the controversial nature of the topic. Learning about a brand's stance that one disapproves of while acknowledging its controversial nature might trigger different reactions than learning about an "objective" moral transgression, such as exploitative practices and child labor would. Moral foundation theory suggests that people share the same universal moral foundations but attach different importance to them; for example, depending on their political beliefs (Graham *et al*., 2009). Therefore, (some) consumers might consider it legitimate to have different opinions on topics such as immigration, gun control, or abortion rights, which makes unaligned brand activism a less severe transgression. In contrast, transgressions such as consumer fraud are universally condemned and, therefore, more severe.

Second, this study adds to a better understanding of the effects of political ideology on consumer responses. The results support previous findings (Ketron *et al*., 2022) and additionally verify the effects when compared to a control group or a brand's intentional "no-stance" message, thus allowing us to test the asymmetric effect of brand activism on liberals and conservatives. Conservatives, in fact, are not as influenced as liberals by activist brands. As expected, liberals respond more negatively to brand activism when it conflicts with their moral values and views. The absence of positive effects—even for liberals—is consistent with the asymmetric effect which was proposed as a general response pattern to brand activism by Mukherjee and Althuizen (2020). In principle, a conservative attitude has a comparable effect with strong brand identification. Both "serve as a safety cushion" (Khamitov *et al.*, 2020, p. 530) in the negative case of consumer–brand disagreement but offer no real advantage when consumers agree with a brand.

Third, this research adds to the emerging literature on perceptions of marginalization and resulting feelings of becoming an "outsider" due to brand actions. So far, this phenomenon was rather restricted to social minority groups that were stigmatized based on their gender or sexual orientation, ethnicity, social class, among others (e.g., Crosby *et al.*, 2015; El-Bassiouny, 2014). Related activities such as CSR are based on consensus in (the majority) society; therefore, such marginalization effects have been not observed on a large scale. In contrast, brand activism is divisive by definition and creates an awareness of a "counterparty" (Hydock *et al.*, 2020), that is, the salience of (not) belonging to certain groups is rising. If the consumers' attention is drawn to these differences through the brand's messaging and they find themselves in opposition to the brand's values, they will feel marginalized. Importantly, this feeling leads to a negative brand evaluation. This finding is highly relevant because feelings of marginalization have been linked to emotional reactions such as anger, fear, frustration, or the desire for retaliation (Betts and Hinsz, 2013). In addition, subjectively marginalized individuals experience a sense of deprivation, a lack of social support and trust in others, and a lack of control over their lives (Bollwerk *et al.*, 2021), which can lead to hostile behavior toward the source of marginalization (Betts and Hinsz, 2013). Perceived marginalization by the brand thus offers an additional explanation for the often extremely negative reactions (e.g., burning Nike shoes) of consumers who disapprove of a brand's stance.

Lastly, the results revealed that conservatives already perceive themselves as more marginalized than liberals, which is probably due to the dominance of liberal brand activism (Klostermann *et al.*, 2021). However, another activist message conservatives disagree with does not increase that feeling but it does for liberals. Further, liberals also feel marginalized by the explicit refusal of brands to take position on a social controversy. The authors believe that this effect can be explained by the importance liberals place on brand activism (Ketron *et al*., 2022). If a company refuses to contribute to social change and to push the liberal agenda, liberals will develop a sense of marginalization and decrease their brand attitude.

Managerial implications

Brand managers in today's politicized world must decide whether and how to get involved in socio-political controversies to drive societal change and stay relevant to their target groups. Even though consumers demand that brands champion sociopolitical issues (Moorman, 2020), the findings emphasize the risks that activism poses to brands. Therefore, this study offers urgently needed managerial guidance on how to engage in activism in a sensible manner. The results provide precise, simple, and easy-to-evaluate prescriptions based on the target group and the topic. Brand managers need to evaluate whether their major customer base has (1) a rather strong or weak CBI and (2) leans to the liberal or conservative side of the political spectrum.

First, when targeting a significant group of weakly identified or liberal consumers, managers must exercise the greatest care not to alienate those consumers. Therefore, brands in specific low-involvement product categories (e.g., commodities) are potentially at significant risk when engaging in brand activism because consumers typically do not feel a strong connection with them and can easily find alternatives. In contrast, brands that have a strong base of highly identified consumers can be more confident in planning activist campaigns.

These customers are found to manifest the "love is blind" effect even when they disagree with the brand's stance—at least in the short term. In the long term, it is imperative for managers to realize that even highly connected consumers will likely become alienated if they are repeatedly confronted with positions they disagree with. Previous research has shown that CBI decreases when consumers do not share the brand's stance (Garg and Saluja, 2022; Mukherjee and Althuizen, 2020), thus continually reducing the "safety cushion". To keep this "safety cushion" as large as possible, managers should focus their marketing efforts (e.g., non-activist communication) on activities that are likely to increase the CBI of their target groups (e.g., see Stokburger-Sauer *et al.*, 2012, Schmalz and Orth, 2012, Swaminathan *et al.*, 2009 and Tuškej *et al.*, 2013 for drivers of CBI). For instance, brands should highlight and develop the brand's uniqueness or prestigious character, strengthen a sense of community and belonging between consumers and the brand, and cultivate positive and distinctive brand associations (Swaminathan *et al.*, 2009). Furthermore, brand managers are advised to create a warm and emotional brand personality and design events that create positive connotations and memorable brand experiences (Stokburger-Sauer *et al.*, 2012; Schmalz and Orth, 2012).

Second, brand managers need to be aware of the high risk involved when taking a clearly conservative position. Liberals tend to have strong negative reactions when they feel their values are attacked by a brand. With liberal stances, which most activist brands have (publicly) shared in the past (Klostermann *et al.*, 2021), brands are more likely to be on the safe side. Therefore, even if a brand's customer base is mainly conservative, liberal stances will probably not cause a severe backlash.

However, this study also shows that conservatives already feel more marginalized than liberals. A single activist statement conservatives disagree with just does not seem to add to that feeling. The authors would encourage managers— independent of the actual stance adopted—to take an inclusive, non-confrontational approach to avoid widening the divide. In particular, managers might use de-escalating elements of communication (e.g., explaining their decision to take a position and avoiding reproachful and offensive messages) (Korschun, 2021). For instance, Burger King used humor as a non-confrontational communication style to support social distancing during the COVID pandemic (i.e., Burger King introduced the social distancing burger with triple onions to keep others away from you) (Campaigns of the World, 2020). As shown by past research, the use of humor can ease tensions and allow difficult topics to be addressed and discussed (Lynch, 2002). Further, conveying positive emotions, such as hope, and showing empathy for both sides could potentially contribute to a constructive atmosphere (e.g., Ahmad *et al*., 2022; Hesse *et al*., 2021; Wannow and Haupt, 2022). Managers also need to understand that companies and their brands have the power to act as moral authorities in society as they can “provide clarity and guidance to others [about] what is right in terms of values and actions” (Hoppner and Vadakkepatt, 2019, p. 418). This power also brings with it the obligation to take responsibility for being a political player when pushing for changes in social values, business practices, or even laws and regulations.

Limitations and future research

This study has several limitations that offer interesting opportunities for future research. First, researchers should continue to unravel the different response patterns to negative brand actions. While this research found a “love is blind” effect, the study should be repeated with alternative topics, for instance, by differentiating between varying degrees of perceived severity or moral intensity from the consumer perspective (Jones, 1991). Understanding the conditions under which the most important customers, namely, the devoted and strongly identified consumers, are particularly tolerant of unaligned brand activism or, on the contrary, extremely sensitive to it, is crucial.

Second, an established and simple measure for political ideology was adopted (for example, see Ketron *et al.*, 2022 and Pecot *et al.*, 2021). A more nuanced analysis of different political ideologies and their strength and extremity is an interesting area for future research.

Because this study only examined brands and consumers in the US market, the findings might be limited to this market. Other countries are characterized by different market conditions and political landscapes. Therefore, further studies on activist brand messaging in different cultural and political contexts are needed.

Third, this study focused on consumers' attitude changes and behavioral intentions (i.e., WoM intentions). Future studies might include real consumer behaviors, such as buycott versus boycott behaviors (e.g., Jungblut and Johnen 2021) and assess further consequences of perceived marginalization. For example, will consumers who feel highly marginalized quietly turn away from the brand or retaliate? Will they still freely express their opinion or remain silent as suggested by the spiral of silence theory? (Hong and Lee, 2021; Noelle-Neumann, 1974). Because a major goal of brand activism is to promote change in society, effects on socially relevant outcomes should be examined. For example, considering the growing polarization, one might explore whether and under which circumstances brand activism adds to or reduces the divide between opposing groups.

Fourth, the studies focused on a one-time activist statement on one specific topic with a neutral message style. Despite the initial, commendable research (Ahmad *et al.*, 2022), more work on specific styles of activist campaigns needs to be done, for instance, by comparing different emotional appeals (e.g., humor, anger, or guilt) or language styles (e.g., neutral, casual, assertive, or even aggressive). In addition, it may be worthwhile to investigate which issues can help brands achieve a more positive impact, e.g., depending on the level of congruency between an issue and a brand (Vredenburg *et al.*, 2020). The evaluation of both long-term approaches to brand activism and their long-term effects would provide valuable insights for brand management. It is possible that brand activism will produce positive outcomes if done consistently and using the "right" message style. Thus, future researchers should attempt to find a way to reduce or even eliminate the asymmetric effect of brand activism (Hydock *et al.*, 2020; Mukherjee and Althuizen, 2020).

5. References

Aggarwal, P. (2004), “The effects of brand relationship norms on consumer attitudes and behavior”, *Journal of Consumer Research*, Vol. 31 No. 1, pp. 87–101.

Ahluwalia, R. (2002), “How prevalent is the negativity effect in consumer environments?”, *Journal of Consumer Research*, Vol. 29 No. 2, pp. 270–279.

Ahluwalia, R., Burnkrant, R.E. and Unnava, H.R. (2000), “Consumer response to negative publicity: the moderating role of commitment”, *Journal of Marketing Research*, Vol. 37 No. 2, pp. 203–214. https://doi.org/10.1509/jmkr.37.2.203.18734

Ahmad, F., Guzmán, F. and Kidwell, B. (2022), “Effective messaging strategies to increase brand love for sociopolitical activist brands, Journal of Business Research, Vol. 151, pp. 609-622. https://doi.org/10.1016/j.jbusres.2022.07.031

Angle, J.W., Dagogo-Jack, S.W., Forehand, M.R. and Perkins, A.W. (2017), “Activating stereotypes with brand imagery: the role of viewer political identity”, *Journal of Consumer Psychology*, Vol. 27 No. 1, pp. 84–90.

Antonetti, P. and Anesa, M. (2017), “Consumer reactions to corporate tax strategies: the role of political ideology”, *Journal of Business Research*, Vol. 74, pp. 1–10. https://doi.org/10.1016/j.jbusres.2016.12.011

Armstrong, J. S., Brodie, R. J. and Parsons, A. G. (2001), “Hypotheses in Marketing Science: Literature Review and Publication Audit”, *Marketing Letters*, Vol. 12 No. 2, pp. 171–187. https://doi.org/10.1023/A:1011169104290

Bailey, J.R. and Phillips, H. (2020), “How Do Consumers Feel When Companies Get Political?”, available at: https://hbr.org/2020/02/how-do-consumers-feel-when-companies-get-political (accessed 29 June 2021).

Baumeister, R. F., Bratslavsky, E., Finkenauer, C. and Vohs, K. D. (2001), “Bad is Stronger than Good”, *Review of General Psychology*, Vol. 5 No. 4, pp. 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Belk, R. W. (1988), “Possessions and the extended self”, *Journal of Consumer Research*, Vol. 15 No. 2, pp. 139–168. https://doi.org/10.1086/209154

Betts, K.R. and Hinsz, V.B. (2013), “Group marginalization: extending research on interpersonal rejection to small groups”, *Personality and Social Psychology Review*, Vol. 17 No. 4, pp. 355–370. https://doi.org/10.1177/1088868313497999

Bhagwat, Y., Warren, N.L., Beck, J.T. and Watson, G.F. (2020), “Corporate sociopolitical activism and firm value”, *Journal of Marketing*, Vol. 84 No. 5, pp. 1–21. https://doi.org/10.1177/0022242920937000

Bhattacharjee, A., Berman, J. Z. and Reed, A. (2013), “Tip of the hat, wag of the finger: How moral decoupling enables consumers to admire and admonish”, *Journal of Consumer Research*, Vol. 39 No. 2, pp. 1167-1184. https://doi.org/10.1086/667786

Bhattacharya, C.B. and Sen, S. (2003), “Consumer-company identification: a framework for understanding consumers’ relationships with companies”, *Journal of Marketing*, Vol. 67 No. 2, pp. 76–88. https://doi.org/10.1509/jmkg.67.2.76.18609

Bollwerk, M., Schlipphak, B. and Back, M.D. (2021), “Development and Validation of the Perceived Societal Marginalization Scale”, *European Journal of Psychological Assessment*, Vol. 38 No. 2, pp. 137-149. https://doi.org/10.1027/1015-5759/a000651

Campaigns of the World. (2020), “Burger King: The Social Distancing Whopper”, available at: https://campaignsoftheworld.com/print/burger-king-the-social-distancing-whopper/ (accessed 28 February 2023)

Chaker, N. N., Walker, D., Nowlin, E. L. and Anaza, N. A. (2019), “When and how does sales manager physical attractiveness impact credibility: A test of two competing hypotheses”, *Journal of Business Research*, Vol. 105, 98-108.

Cheng, S.Y.Y., White, T.B. and Chaplin, L.N. (2012), “The effects of self-brand connections on responses to brand failure: a new look at the consumer–brand relationship”, *Journal of Consumer Psychology*, Vol. 22 No. 2, pp. 280–288.

Copeland, L. and Boulianne, S. (2022), “Political consumerism: a meta-analysis”, *International Political Science Review*, Vol. 43 No. 1, pp. 3–18.

Crosby, E., McKeage, K. and Rittenburg, T. (2015), “Marginalization, resistance, and stealth shopping: subalternity in the gendered marketplace”, in Diehl, Kristin and Yoon, C. (Eds.), *NA - Advances in Consumer Research*, 43rd ed., Duluh, MN, pp. 254–259.

Dalman, M.D., Buche, M.W. and Min, J. (2019), “The differential influence of identification on ethical judgment: the role of brand love”, *Journal of Business Ethics*, Vol. 158 No. 3, pp. 875–891.

Davidson, A. and Theriault, D.A. (2021), “How Consumer Experience Is Shaped by the Political Orientation of Service Providers”, *Journal of Consumer Psychology*, Vol. 31 No. 4, pp. 792–800. https://doi.org/10.1002/jcpy.1233

Davvetas, V. and Diamantopoulos, A. (2017), "“Regretting your brand-self?” The moderating role of consumer-brand identification on consumer responses to purchase regret”, *Journal of Business Research*, Vol. 80, pp. 218–227. https://doi.org/10.1016/j.jbusres.2017.04.008

Day, M.V., Fiske, S.T., Downing, E.L. and Trail, T.E. (2014), "Shifting liberal and conservative attitudes using moral foundations theory", *Personality and Social Psychology Bulletin*, Vol. 40 No. 12, pp. 1559–1573.

Duman, S. and Ozgen, O. (2018), "Willingness to punish and reward brands associated to a political ideology (BAPI)", *Journal of Business Research*, Vol. 86, pp. 468–478. https://doi.org/10.1016/j.jbusres.2017.05.026

Einwiller, S., Fedorikhin, A., Johnson, A.R. and Kamins, M.A. (2006), "Enough is enough! When identification no longer prevents negative corporate associations", *Journal of the Academy of Marketing Science*, Vol. 34 No. 2, pp. 185–194. https://doi.org/10.1177/0092070305284983

Einwiller, S., Lis, B., Ruppel, C. and Sen, S. (2019), "When CSR-based identification backfires: testing the effects of CSR-related negative publicity", *Journal of Business Research*, Vol. 104, pp. 1–13.

El-Bassiouny, N. (2014), "The one-billion-plus marginalization: toward a scholarly understanding of Islamic consumers", *Journal of Business Research*, Vol. 67 No. 2, pp. 42–49. https://doi.org/10.1016/j.jbusres.2013.03.010

Escalas, J.E. and Bettman, J.R. (2005), "Self-construal, reference groups, and brand meaning", *Journal of Consumer Research*, Vol. 32 No. 3, pp. 378–389. https://doi.org/10.1086/497549

Fernandes, D. (2020), "Politics at the mall: the moral foundations of boycotts", *Journal of Public Policy & Marketing*, Vol. 39 No. 4, pp. 494–513. https://doi.org/10.1177/0743915620943178

Folkes, V. S. and Kamins, M. A. (1999), „Effects of information about firms' ethical and unethical actions on consumers' attitudes", *Journal of Consumer Psychology*, Vol. 8 No. 3, pp. 243-259.

Gallup, Inc. (2022), "Abortion", available at: https://news.gallup.com/poll/1576/Abortion.aspx (accessed 21 July 2022)

Gardberg, N.A. and Newburry, W. (2013), "Who Boycotts Whom? Marginalization, Company Knowledge, and Strategic Issues", *Business & Society*, Vol. 52 No. 2, pp. 318–357. https://doi.org/10.1177/0007650309352507

Garg, N. and Saluja, G. (2022), "A tale of two "ideologies": differences in consumer response to brand activism", *Journal of the Association for Consumer Research*, Vol. 7 No. 3, p. 325-339 .

Goldberg, E. (2022), "These Companies Will Cover Travel Expenses for Employee Abortions", *The New York Times,* 27 June, available at: https://www.nytimes.com/article/abortion-companies-travel-expenses.html (accessed 21 July 2022)

Graham, J., Haidt, J. and Nosek, B. A. (2009), "Liberals and conservatives rely on different sets of moral foundations", *Journal of personality and social psychology*, Vol. 96 No. 5, 1029-1046. https://doi.org/10.1037/a0015141

Grégoire, Y. and Fisher, R.J. (2006), "The effects of relationship quality on customer retaliation", *Marketing Letters*, Vol. 17 No. 1, pp. 31–46.

Grégoire, Y. and Fisher, R.J. (2008), "Customer betrayal and retaliation: when your best customers become your worst enemies", *Journal of the Academy of Marketing Science*, Vol. 36 No. 2, pp. 247–261.

Grégoire, Y., Tripp, T.M. and Legoux, R. (2009), "When customer love turns into lasting hate: the effects of relationship strength and time on customer revenge and avoidance", *Journal of Marketing*, Vol. 73 No. 6, pp. 18–32.

Hall, J. M. (1999), "Marginalization revisited: Critical, postmodern, and liberation perspectives", *Advances in Nursing Science*, Vol. 22 No. 2, pp. 88–102. https://doi.org/10.1097/00012272-199912000-00009

Hambrick, D.C. and Wowak, A.J. (2021), "CEO sociopolitical activism: a stakeholder alignment model", *Academy of Management Review*, Vol. 46 No. 1, pp. 33–59. https://doi.org/10.5465/amr.2018.0084

Hayes, A.F. (2021), *Introduction to mediation, moderation, and conditional process analysis: a regression-based approach,* Third edition, Guilford publications, New York, NY.

He, H. and Li, Y. (2011), "CSR and service brand: The mediating effect of brand identification and moderating effect of service quality", *Journal of Business Ethics*, Vol. 100 No. 4, pp. 673–688.

Hesse, A., Niederle, F., Schön, L. and Stautz, D. (2021), "Consumer responses to brand communications involving COVID-19", *Journal of Marketing Management*, Vol. 37 No.17-18, pp. 1783–1814. https://doi.org/10.1080/0267257X.2021.2000008

Hong, C. and Li, C. (2021), "Will consumers silence themselves when brands speak up about sociopolitical issues? Applying the spiral of silence theory to consumer boycott and buycott behaviors", *Journal of nonprofit & public sector marketing*, Vol. 33 No, 2, pp. 193-211.

Hoppner, J.J. and Vadakkepatt, G.G. (2019), "Examining moral authority in the marketplace: a conceptualization and framework", *Journal of Business Research*, Vol. 95, pp. 417–427.

Hoyer, Wayne D.; MacInnis, Deborah J.; Pieters, Rik (2018): *Consumer behavior*, 7th edition, Boston, MA: Cengage Learning.

Hulland, J., Baumgartner, H. and Smith, K.M. (2018), "Marketing survey research best practices: evidence and recommendations from a review of JAMS articles", *Journal of the Academy of Marketing Science*, Vol. 46 No. 1, pp. 92–108. https://doi.org/10.1007/s11747-017-0532-y

Hydock, C., Paharia, N. and Blair, S. (2020), "Should your brand pick a side? How market share determines the impact of corporate political advocacy", *Journal of Marketing Research*, Vol. 57 No. 6, pp. 1135–1151. https://doi.org/10.1177/0022243720947682

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. and Westwood, S.J. (2019), "The origins and consequences of affective polarization in the United States", *Annual Review of Political Science*, Vol. 22, pp. 129–146.

Jain, K. and Sharma, I. (2019), "Negative outcomes of positive brand relationships", *Journal of Consumer Marketing*, Vol. 36 No. 7, pp. 986–1002. https://doi.org/10.1108/JCM-07-2018-2764

Jetten, J., Branscombe, N.R., Schmitt, M.T. and Spears, R. (2001), "Rebels with a cause: group identification as a response to perceived discrimination from the mainstream", *Personality and Social Psychology Bulletin*, Vol. 27 No. 9, pp. 1204–1213.

Jones, T.M. (1991), "Ethical decision making by individuals in organizations: an issue-contingent model", *Academy of Management Review*, Vol. 16 No. 2, pp. 366–395. https://doi.org/10.5465/amr.1991.4278958

Jost, J.T., Federico, C.M. and Napier, J.L. (2009), "Political ideology: its structure, functions, and elective affinities", *Annu. Rev. Psychol.*, Vol. 60 No. 1, pp. 307–337. https://doi.org/10.1146/annurev.psych.60.110707.163600

Jost, J.T., Nosek, B.A. and Gosling, S.D. (2008), "Ideology: its resurgence in social, personality, and political psychology", *Perspectives on Psychological Science*, Vol. 3 No. 2, pp. 126–136.

Jung, J. and Mittal, V. (2020), "Political identity and the consumer journey: a research review", *Journal of Retailing*, Vol. 96 No. 1, pp. 55–73. https://doi.org/10.1016/j.jretai.2019.09.003

Jung, K., Garbarino, E., Briley, D.A. and Wynhausen, J. (2017), "Blue and Red Voices: Effects of Political Ideology on Consumers' Complaining and Disputing Behavior", *Journal of Consumer Research*, Vol. 44 No. 3, pp. 477–499.

Jungblut, M. and Johnen, M. (2021), "When brands (don't) take my stance: the ambiguous effectiveness of political brand communication", *Communication research*, pp. 1-26. https://doi.org/10.1177/00936502211001622

Kaplan, A. M., Schoder, D. and Haenlein, M. (2007), "Factors Influencing the Adoption of Mass Customization: The Impact of Base Category Consumption Frequency and Need Satisfaction", *Journal of Product Innovation Management*, Vol. 24 No. 2, pp.101–116. https://doi.org/10.1111/j.1540-5885.2007.00237.x

Ketron, S., Kwaramba, S. and Williams, M. (2022), "The "company politics" of social stances: how conservative vs. liberal consumers respond to corporate political stance-taking", *Journal of Business Research*, Vol. 146, pp. 354–362. https://doi.org/10.1016/j.jbusres.2022.03.086

Key, T.M., Keel, A.L., Czaplewski, A.J. and Olson, E.M. (2021), "Brand activism change agents: strategic storytelling for impact and authenticity", *Journal of Strategic Marketing*, pp. 1–17. https://doi.org/10.1080/0965254X.2021.1904435

Khamitov, M., Grégoire, Y. and Suri, A. (2020), "A systematic review of brand transgression, service failure recovery and product-harm crisis: Integration and guiding insights", *Journal of the Academy of Marketing Science*, Vol. 48 No. 3, pp. 519–542. https://doi.org/10.1007/s11747-019-00679-1

Kim, T.J., Yi, Y. and Choi, J. (2020), "The boomerang effect of brand personality congruency in a product-harm crisis", *Australian Journal of Management*, Vol. 45 No. 4, pp. 645–661. https://doi.org/10.1177/0312896219895059

Klostermann, J., Hydock, C. and Decker, R. (2021), "The effect of corporate political advocacy on brand perception: an event study analysis", *Journal of Product & Brand Management,* Vol. 31 No. 5, pp. 780-797. https://doi.org/10.1108/JPBM-03-2021-3404

Korschun, D. (2021), "Brand Activism Is Here to Stay: Here's Why". *NIM Marketing Intelligence Review*, Vol. 13 No. 2, pp. 10–17. https://doi.org/10.2478/nimmir-2021-0011

Korschun, D., Martin, K.D. and Vadakkepatt, G. (2020), “Marketing’s role in understanding political activity”, *Journal of Public Policy & Marketing*, Vol. 39 No. 4, pp. 378–387. https://doi.org/10.1177/0743915620949261

Kuenzel, S. and Vaux Halliday, S. (2008), “Investigating antecedents and consequences of brand identification”, *Journal of Product & Brand Management*, Vol. 17 No. 5, pp. 293–304. https://doi.org/10.1108/10610420810896059

Lam, S.K., Ahearne, M., Hu, Y. and Schillewaert, N. (2010), “Resistance to brand switching when a radically new brand is introduced: a social identity theory perspective”, *Journal of Marketing*, Vol. 74 No. 6, pp. 128–146. https://doi.org/10.1509/jmkg.74.6.128

Leak, R.L., Woodham, O.P. and McNeil, K.R. (2015), “Speaking candidly: how managers’ political stances affect consumers’ brand attitudes”, *Journal of Product & Brand Management*, Vol. 24 No. 5, pp. 494–503. https://doi.org/10.1108/JPBM-08-2014-0693

Lee, H., Chang, D. R. and Einwiller, S. (2020), „ A study on the dynamics between the moral reasoning process and celebrity image and their impact on consumers’ support for celebrity comebacks after a transgression”, *Journal of Product & Brand Management*, Vol. 29 No. 6, pp. 729-743. 10.1108/JPBM-02-2019-2259

Liu, T.-C., Wang, C.-Y. and Wu, L.-W. (2010), “Moderators of the negativity effect: commitment, identification, and consumer sensitivity to corporate social performance”, *Psychology & Marketing*, Vol. 27 No. 1, pp. 54–70.

Lynch, O. H. (2002), “Humorous communication: Finding a place for humor in communication research”, *Communication theory*, Vol. 12 No. 4, pp. 423-445.

Ma, L. (2020), “When love becomes hate: how different consumer-brand relationships interact with crises to influence consumers’ reactions”, *Corporate Communications: An International Journal*, Vol. 25 No. 3, pp. 357–375. https://doi.org/10.1108/CCIJ-08-2019-0103

Mady, T. T. (2011), “Sentiment toward marketing: Should we care about consumer alienation and readiness to use technology?”, *Journal of Consumer Behaviour*, Vol. 10 No. 4, pp. 192–204. https://doi.org/10.1002/cb.329

Mason, L. (2015), ““I disrespectfully agree”: the differential effects of partisan sorting on social and issue polarization”, *American Journal of Political Science*, Vol. 59 No. 1, pp. 128–145.

Matos, G., Vinuales, G. and Sheinin, D.A. (2017), “The power of politics in branding”, *Journal of marketing theory and practice*, Vol. 25 No. 2, pp. 125–140. https://doi.org/10.1080/10696679.2016.1270768

Mirzaei, A., Wilkie, D.C. and Siuki, H. (2022), “Woke brand activism authenticity or the lack of it”, *Journal of Business Research*, Vol. 139, pp. 1–12.

Moorman, C. (2020), “Commentary: brand activism in a political world”, *Journal of Public Policy & Marketing*, Vol. 39 No. 4, pp. 388–392. https://doi.org/10.1177/0743915620945260

Mukherjee, S. and Althuizen, N. (2020), “Brand activism: does courting controversy help or hurt a brand?”, *International Journal of Research in Marketing*, Vol. 37 No. 4, pp. 772–788. https://doi.org/10.1016/j.ijresmar.2020.02.008

Napier, J.L. and Jost, J.T. (2008), “Why are conservatives happier than liberals?”, *Psychological Science*, Vol. 19 No. 6, pp. 565–572.

Noelle-Neumann, E. (1974), “The spiral of silence a theory of public opinion”, *Journal of communication*, Vol. 24 No. 2, pp. 43-51.

Northey, G., Dolan, R., Etheridge, J., Septianto, F. and van Esch, P. (2020), “LGBTQ Imagery in Advertising”, Journal of Advertising Research, Vol. 60 No. 2, pp. 222–236. https://doi.org/10.2501/JAR-2020-009

Papaoikonomou, E., Cascon-Pereira, R. and Ryan, G. (2016), “Constructing and communicating an ethical consumer identity: a Social Identity Approach”, *Journal of Consumer Culture*, Vol. 16 No. 1, pp. 209–231. https://doi.org/10.1177/1469540514521080

Pecot, F., Vasilopoulou, S. and Cavallaro, M. (2021), “How political ideology drives anti-consumption manifestations”, *Journal of Business Research*, Vol. 128, pp. 61–69.

Pullig, C., Netemeyer, R. G. and Biswas, A. (2006), “Attitude basis, certainty, and challenge alignment: A case of negative brand publicity”, *Journal of the Academy of Marketing Science*, Vol. 34 No. 4, pp. 528–542. https://doi.org/10.1177/0092070306287128

Rasmussen, K., Dufur, M.J., Cope, M.R. and Pierce, H. (2021), “Gender marginalization in sports participation through advertising: the case of Nike”, *International Journal of Environmental Research and Public Health*, Vol. 18 No. 15, pp. 1–22. https://doi.org/10.3390/ijerph18157759

Saavedra Torres, J.L., Rawal, M. and Bagherzadeh, R. (2020), “Role of brand attachment in customers’ evaluation of service failure”, *Journal of Product & Brand Management*, Vol. 30 No. 3, pp. 377–391. https://doi.org/10.1108/JPBM-03-2019-2293

Sawyer, A. G. and Peter, J. P. (1983), “The Significance of Statistical Significance Tests in Marketing Research”, *Journal of Marketing Research*, Vol. 20 No. 2, pp. 122–133. https://doi.org/10.1177/002224378302000203

Schmalz, S. and Orth, U.R. (2012), “Brand attachment and consumer emotional response to unethical firm behavior”, *Psychology & Marketing*, Vol. 29 No. 11, pp. 869–884.

Schmidt, H.J., Ind, N., Guzmán, F. and Kennedy, E. (2021), “Sociopolitical activist brands”, *Journal of Product & Brand Management*, No. 1 Vol. 1, pp. 40-55. https://doi.org/10.1108/JPBM-03-2020-2805

Sen, S. and Bhattacharya, C. B. (2001), “Does Doing Good Always Lead to Doing Better? Consumer Reactions to Corporate Social Responsibility”, *Journal of Marketing Research*, Vol. 38 No. 2, pp. 225–243. https://doi.org/10.1509/jmkr.38.2.225.18838

Sonnenfeld, J., Tian, S. and Hirsty, G. (2022), “A list of companies supporting abortion rights after the Roe v. Wade ruling shows which firms are stepping up, and why”, *Fortune,* 30 June, available at: https://fortune.com/2022/06/30/companies-supporting-abortion-rights-roe-v-wade-first-movers/ (accessed 21 July 2022)

Spiller, S. A., Fitzsimons, G. J., Lynch, J. G. and Mcclelland, G. H. (2013), “Spotlights, Floodlights, and the Magic Number Zero: Simple Effects Tests in Moderated Regression”, *Journal of Marketing Research*, Vol. 50 No. 2, pp. 277–288. https://doi.org/10.1509/jmr.12.0420

Stokburger-Sauer, N., Ratneshwar, S. and Sen, S. (2012), “Drivers of consumer-brand identification”, *International Journal of Research in Marketing*, Vol. 29 No. 4, pp. 406–418. https://doi.org/10.1016/j.ijresmar.2012.06.001

Strohminger, N. and Nichols, S. (2014), “The essential moral self”, *Cognition*, Vol. 131 No. 1, pp. 159-171.

Swaminathan, V., Stilley, K. M. and Ahluwalia, R. (2009), “When Brand Personality Matters: The Moderating Role of Attachment Styles”, *Journal of Consumer Research*, Vol. 35 No. 6, pp. 985–1002. https://doi.org/10.1086/593948

Tajfel, H. and Turner, J.C. (1979), “An integrative theory of intergroup conflict”, in Austin, W.G. and Worchel, S. (Eds.), *The social psychology of intergroup relations*, Brooks/Cole Publishing Company, Montierey, CA, pp. 33-47.

Taylor, S. E. (1991), “Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis”, *Psychological Bulletin,* Vol. 110 No. 1, pp. 67–85. https://doi.org/10.1037/0033-2909.110.1.67

Trudel, R. and Cotte, J. (2009), “Does It Pay To Be Good?”, *MIT Sloan Management Review*, available at: https://sloanreview.mit.edu/article/does-it-pay-to-be-good/ (accessed 28 February 2023)

Trump, R.K. (2014), “Connected consumers’ responses to negative brand actions: The roles of transgression self-relevance and domain”, *Journal of Business Research*, Vol. 67 No. 9, pp. 1824–1830. https://doi.org/10.1016/j.jbusres.2013.12.007

Tuškej, U., Golob, U. and Podnar, K. (2013), “The role of consumer–brand identification in building brand relationships”, *Journal of Business Research*, Vol. 66 No. 1, pp. 53–59.

Vredenburg, J., Kapitan, S., Spry, A. and Kemper, J.A. (2020), “Brands taking a stand: authentic brand activism or woke washing?”, *Journal of Public Policy & Marketing*, Vol. 39 No. 4, pp. 444–460. https://doi.org/10.1177/0743915620947359

Wan, L.C., Hui, M.K. and Wyer Jr, R.S. (2011), “The role of relationship norms in responses to service failures”, *Journal of Consumer Research*, Vol. 38 No. 2, pp. 260–277.

Wannow, S. and Haupt, M. (2022), “When Brands Take a Stand: Navigating Emotional Reactions to Brand Activism”, *Marketing Review St. Gallen*, Vol. 39 No.2, pp. 44–51.

Weber, T.J., Joireman, J., Sprott, D.E. and Hydock, C. (2023), “Differential Response to Corporate Political Advocacy and Corporate Social Responsibility: Implications for Political Polarization and Radicalization”, Journal of Public Policy & Marketing, Vol. 42 No. 1, pp. 74–93. https://doi.org/10.1177/07439156221133073

Weitzl, W.J., Hutzinger, C. and Wagner, U. (2023), “I am ashamed of my brand-self! Consumer-brand identification as a moderator of emotional reactions following symbol-laden brand failures”, Journal of Product & Brand Management, Vol. ahead-of-print No. ahead-of print. https://doi.org/10.1108/JPBM-02-2022-3853

White, K. and Argo, J.J. (2009), “Social identity threat and consumer preferences”, *Journal of Consumer Psychology*, Vol. 19 No. 3, pp. 313–325.

Xiao, N. and Hwan Lee, S. (2014), “Brand identity fit in co-branding”, *European Journal of Marketing*, 48(7/8), pp. 1239–1254. https://doi.org/10.1108/EJM-02-2012-0075

Xie, C., Bagozzi, R. P. and Grønhaug, K. (2019), “The impact of corporate social responsibility on consumer brand advocacy: The role of moral emotions, attitudes, and individual differences”, *Journal of Business Research*, Vol. 95, pp. 514–530.

Zhang, Y., Zhang, J. and Sakulsinlapakorn, K. (2020), “Love becomes hate? Or love is blind? Moderating effects of brand love upon consumers’ retaliation towards brand failure”, *Journal of Product & Brand Management*, Vol. 30 No. 3, pp. 415–432. https://doi.org/10.1108/JPBM-07-2019-2470

Figure 1. Conceptual model

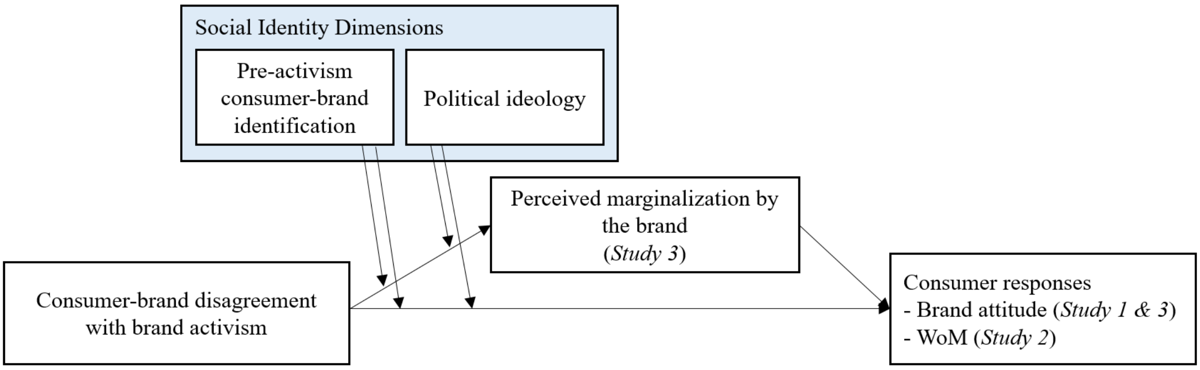
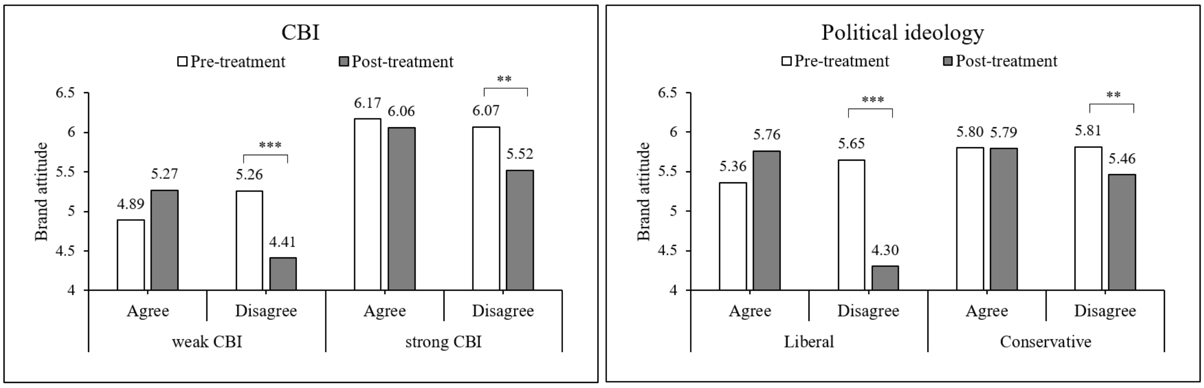
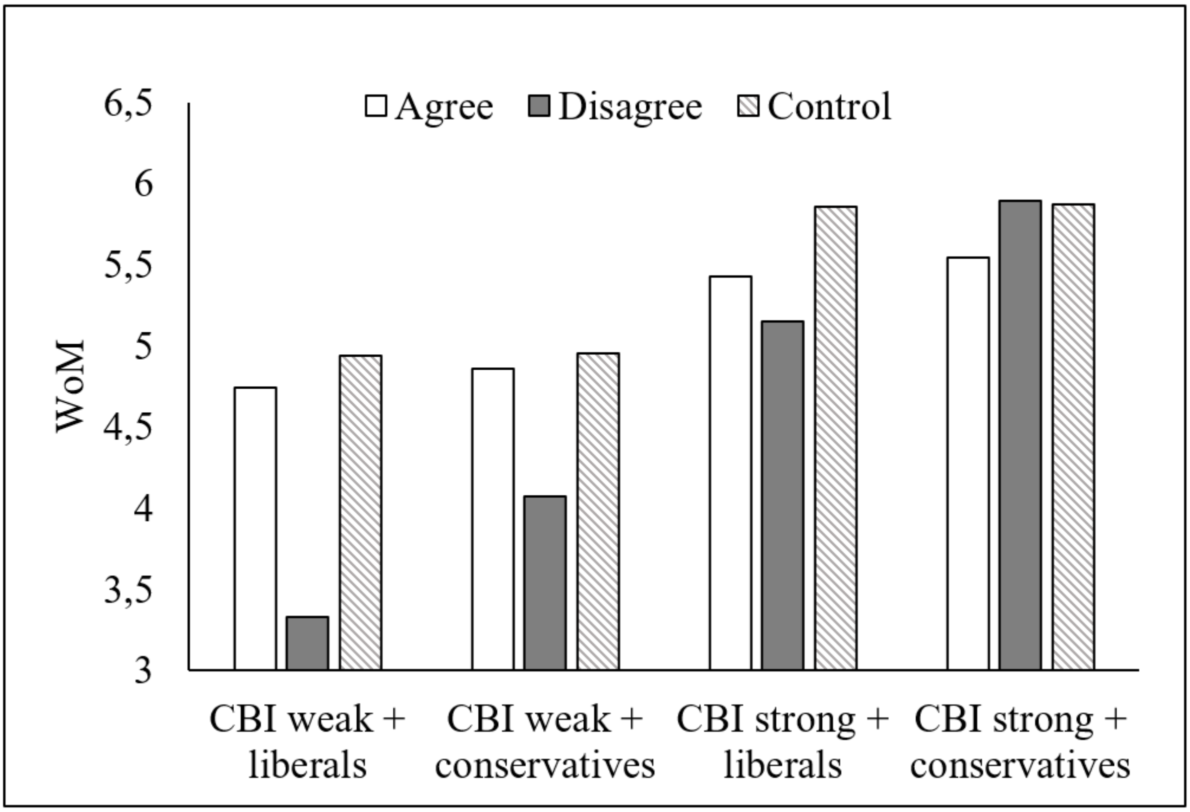

Social Identity Dimensions
Pre-activism consumer-brand identification
Political ideology
Perceived marginalization by the brand (*Study 3*)
Consumer-brand disagreement with brand activism
Consumer responses
- Brand attitude (*Study 1 & 3*)
- WoM (*Study 2*)

Figure 2 Estimated marginal means related to CBI and political ideology depending on consumer-brand agreement versus disagreement (Study 1)



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

Figure 3 Results of spotlight analyses (Study 2)



Note: Values represent estimated conditional means. CBI weak (-1 SD = 3.35) versus CBI strong (+1 SD = 6.49); political orientation: liberals (-1 SD = 2.50) versus conservatives (+1 SD = 6.56).

Figure 4 Results from Study 3

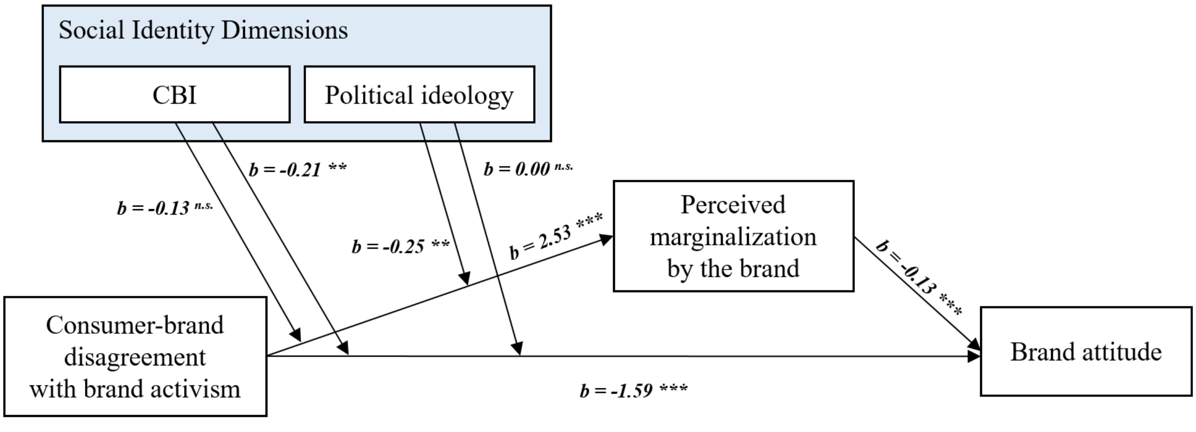


Table 1 Scale items and statistics

Construct name and items	Standardized loadings		
	Study 1	Study 2	Study 3
Consumer–brand identification (CBI)			
(Study 1/2/3: α = .94/.94/.96; AVE = .75/.72/.78; CR = .93/.93/.95)			
• I feel a strong sense of belonging to the brand.	.89	.84	.87
• I identify strongly with the brand.	.87	.85	.89
• The brand embodies what I believe in.	.80	.84	.88
• The brand is like a part of me.	.88	.82	.89
• The brand has a great deal of personal meaning for me.	.88	.87	.90
Perceived marginalization by the brand			
(Study 3: α = .94; AVE = .79; CR = .95)			
• The brand does not care what people like me think.			.89
• The problems of people like me are unimportant to this brand.			.90
• The concerns of people like me are not noticed by this brand.			.90
• The problems of people like me are not heard by this brand.			.91
• This brand pays too little attention to the interests of people like me.			.84
Political ideology			
(Study 1/2/3: α = not applicable, AVE = .94/.94/.91; CR = .94/.94/.91)			
Please state your political orientation (1= very liberal to 7 = very conservative)	.97	.97	.96
Brand attitude			
(Study 1/3: α = .94/.95; AVE = .82/.70; CR = .93/.88)			
• Bad – Good	.89		.83
• Unpleasant – Pleasant	.91		.82
• Dislike – Like	.92		.86
Word-of-mouth (WoM)			
(Study 2: α = .91; AVE = .74; CR = .90)			
Please describe how you would talk to friends, relatives, and other people about [the brand]. I intend to …			
• … talk 1 – negatively to 7 – positively.		.89	
• … recommend this company as a place to work by 1 – no means to 7 – all means		.80	
• … 1 – discredit the company to 7 – speak well of the company.		.90	

Table 2. Repeated measures ANCOVA for brand attitude (Study 1)

Model for CBI as moderator

Factor	df	*F*	$\acute{\eta}^2$	*P*
Between-subjects	222	(2.02)		
Disagreement (DIS)	1	4.32*	0.02	0.54
CBI	1	49.94***	0.18	1.00
DIS x CBI	1	0.09	0.00	0.06
COV: Political ideology (PI)	1	19.08***	0.08	0.99
Within-subjects	222	(1.04)		
Brand attitude (BAT)	1	9.97**	0.04	0.88
BAT x DIS	1	18.17***	0.08	0.99
BAT x CBI	1	0.23	0.01	0.08
BAT x DIS x CBI	1	4.20*	0.02	0.53
BAT x PI	1	4.49*	0.02	0.56

Model for political ideology as moderator

Factor	df	*F*	$\acute{\eta}^2$	*P*
Between-subjects	201	(1.47)		
Disagreement (DIS)	1	8.57**	0.04	0.83
Political ideology (PI)	1	11.16***	0.05	0.91
DIS x PI	1	2.73	0.01	0.38
COV: CBI	1	127.65***	0.39	1.00
Within-subjects	201	(1.05)		
Brand attitude (BAT)	1	0.40	0.00	0.10
BAT x DIS	1	23.43***	0.10	1.00
BAT x PI	1	1.73	0.01	0.26
BAT x DIS x PI	1	10.37**	0.05	0.89
BAT x CBI	1	0.08	0.00	0.06

Note: * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$; $\acute{\eta}^2$ = Partial Eta Squared*,* *P* = Observed Power, values enclosed in parentheses represent mean square errors.

Appendix 1. Literature Review – Role of Consumer-brand identification (CBI) in Controversial or Negative Brand Events

Source	**Topic**	**Focal Relationship Variable(s)**	**Mediator(s), Moderator(s), Dependent Variable(s)**	**Main Findings Regarding Role of CBI**
Aggarwal, 2004	Brand actions in "request for help"-situations (charging a fee for a special service; type of compensation for survey participation)	Relationship type/norms: "exchange" vs. "communal"	Med: Perceived norm violation DV: Reactions to marketing actions, brand evaluation	Consistency between relationship norms and brand actions determines consumer evaluations. E.g., charging a fee for a special service leads to poor evaluations from consumers with communal norms, but not from those with exchange norms. →**Amplifier effect** when relationship norms are violated
Ahluwalia *et al.*, 2000	Positive vs. negative product-related publicity	Commitment	Med: Perceived diagnosticity, counterarguments DV: Attitude ambivalence, attitude change	Brand attitudes of committed consumers are less affected by negative information. Less committed consumers rate the diagnosticity of negative information higher, counter argue it to a lesser degree, and exhibit greater attitude change and ambivalence. → **Buffer effect**
Antonetti and Anesa, 2017	Aggressive vs. conservative corporate tax strategies (CTSs)	Self-brand connection (favorite vs. least favorite brands)	Med: Perceived ethicality of the firm Mod: individuals' political identification DV: Attitudes toward the corporation, negative WoM, purchase intention	Consumer reactions to CTSs depends on the customer's personal connection to the brand: Connected consumers punish brands engaging in CTSs more than less connected consumers. →**Amplifier effect**
Cheng, White and Chaplin, 2012	Negative brand information	Self-brand connections	Mod: Self-brand connections, self-affirmation DV: Brand evaluation	Higher self-brand connection buffers effect of negative brand information, but when high SBC consumers could confirm themselves the negative effect is stronger → **Buffer effect** when self-affirmation is not given

Dalman *et al.*, 2019	Unethical brand situations	CBI and brand love	Med: Ethical judgement, brand love DV: Sin of omission, brand defense	Brand identification both decreases (direct effect) and increases (indirect effect through brand love) consumers' ethical judgment following extremely unethical events. However, the total effect of brand identification on ethical judgement is positive. → **Buffer effect**
Davvetas and Diamantopoulos, 2017	Purchase regret	CBI	Med: Satisfaction, DV: Behavioral intentions (brand repurchase and recommendation)	Consumer-brand identification reduces the negative effects of purchase regret on behavioral intentions both directly and indirectly via satisfaction → **Buffer effect**
Einwiller *et al.*, 2006	Moderately vs. extremely negative publicity	Consumer-company identification (CCI)	DV: Behavioral intentions and attitude toward the company	CCI only decreases negative effects on behavioral intentions and attitudes for moderately negative publicity. No difference between strong and weak identifiers' reactions for extremely negative publicity → **Buffer effect** for moderately negative events
Einwiller *et al.*, 2019	Negative publicity based on ethical misconduct (CSR-related) vs. lack of corporate ability	Consumer-company identification (CCI)	Med: Perceived betrayal, consumer disidentification DV: Attitude towards the company, intention to brand opposition	CCI buffers negative effects of negative publicity in case of negative publicity related to corporate ability. For CSR-related negative publicity the effect is reversed. → **Buffer effect** for negative events related to corporate ability →**Amplifier effect** for negative events related to CSR
Grégoire and Fisher, 2006	High vs. low controllable service failures	Relationship quality (RQ)	Med: Desire for retaliation DV: Negative WoM, third party complaining, patronage reduction	Under low controllability conditions, high RQ reduces desire for retaliation. Under high controllability conditions, the desire for retaliation does not significantly differ between high vs. low RQ customers. → **Buffer effect** for service failures when low controllability is inferred
Grégoire and Fisher, 2008	Unresolved service failures	Relationship quality (RQ)	Med: Perceived betrayal	High (vs. low) RQ customers feel more betrayed in case of violation of distributive and process fairness. → **Amplifier effect** in case of severe fairness violations

			DV: Retaliatory behavior, demands for reparation	
Grégoire *et al.*, 2009	Online-complaining context	Relationship quality (RQ)	Med: Perceived betrayal DV: Desire for revenge, avoidance	After their online complaints, high- (vs. low-) RQ customers keep their desire for revenge for longer. Their desire for avoidance also increases more rapidly over time. → **Amplifier effect**, i.e., longitudinal love-becomes-hate effect for online complainers
Jain and Sharma, 2019	Product failure	Brand attachment	Med: External attribution of controllability, perceived betrayal, active and passive brand hate , Mod: Narcissistic entitlement DV: eWOM on social media	High (vs. no) brand attachment drives feelings of betrayal in case of a product failure – directly and indirectly via external attribution – leading to higher levels of brand hate and negative eWoM. →**Amplifier effect**
Jeon and Baeck, 2016	Negative CSR and product-related information	Brand-customer relationship strength	DV: Attitudes, Behavioral intentions	Stronger relationships and CSR (product-related) information reduce negative effect of negative CSR (product-related) information → **Buffer effect**
Kim *et al.*, 2020	Product-harm crisis	Personality congruency	Med: Disappointment, locus, stability, controllability, blame DV: Purchase intention	Personality congruency increases the negative influence of a product-harm crisis on purchase intentions via increased disappointment, failure attributions to the brand. →**Amplifier effect**
Leak *et al.*, 2015	Managers' political stances regarding same-sex marriage	Pre-existing brand attitudes	Med: Anger DV: Brand attitude change	Preexisting brand attitude leads to positive attitude change after exposure to a managers' stance – directly and indirectly vie reduced anger → **Buffer effect** in case of disagreement with managers' stances

Liu *et al.*, 2010	Moderately vs. extremely negative/positive publicity	Customer commitment, consumer-company identification (CCI)	Mod: Customer sensitivity to corporate social performance , DV: Attitude change	Commitment and CCI moderate the effects of moderately negative (vs. positive) publicity on attitude change toward a retailer. No significant moderation effect for extremely negative information. → **Buffer effect** only for moderately negative events
Ma, 2020	Brand transgressions	Consumer-brand relationships	Med: Attitude, anger, disappointment, sympathy DV: negative WoM intentions, purchase intentions	Non-identifying relationships act like a buffer against brand transgressions. Identifying relationships primarily lead to a love-becomes-hate effect by amplifying negative emotions and disappointment and, thereby, reducing behavioral intentions – regardless of brand transgression type. → **Buffer** and **amplifier effect** depending on type of relationship
Pullig, Netemeyer and Biswas, 2006	Negative publicity	Prior attitude certainty	DV: Attitude change	Stronger prior attitude certainty reduces effect of negative publicity when brand positioning is aligned (vs. not) with negative brand publicity (performance- or value-based information). → **Buffer effect** when brand attitude certainty is high
Saavedra Torres *et al.*, 2020	Influence of brand attachment on the effect of a **service failure** (i.e., stability and control attributions) on customers' negative emotions	Brand attachment	Med: Negative emotions DV: WoM, loyalty intentions	Brand attachment can reduce customers' negative emotions – especially in cases of a controllable service failure → **Buffer effect** when service failure is attributed to a controllable cause
Schmalz and Orth, 2012	Unethical firm behavior	Brand attachment	Med: Ethical judgment, emotional ambivalence, DV: Purchase intention	Attachment attenuates judgments of unethical behavior, contributes to emotional ambivalence, and affects purchase intentions – but only for moderately unethical firm behavior. → **Buffer effect** only for moderately negative events

Trump, 2014	Negative brand actions	Consumer-brand connection	Mod: Self-relevance DV: Brand evaluation	Brand evaluations of strongly connected consumers decrease in response to self-relevant product failures or unethical behavior. → **Buffer effect** only for product-related brand transgressions that are not self-relevant
Wan *et al.*, 2011	Service failures	Relationship type/norms: “exchange” vs. “communal”	Med: Feelings of betrayal, self- vs. other obligation thoughts, attribution of responsibility DV: Negative reaction to service failure	Moderating effect of communal norms can be negative or positive depending on the specific circumstances. →**Amplifier effect** of close relationship when (a) consumers think about their service provider’s obligations in the relationship, (b) a promised service was not provided, and (c) independent self-construal is active.
Weitzl *et al.*, 2023	Dissatisfaction after self-relevant, symbolic brand failures	CBI	Med: Brand shame DV: Brand anger	CBI intensifies the direct and indirect (via brand shame) negative effect of dissatisfaction after self-relevant, symbolic brand failures on brand anger. → **Amplifier effect**
Zhang *et al.*, 2020	Brand failures	Brand love	Mod: Perceived fairness, inferred goodwill, aggressive personality, brand trust Med: Negative emotions DV: Consumer retaliation	Brand love amplifies the effect of failure severity on negative emotions (‘love becomes hate’). At the same time, brand love weakens the relationship between negative emotions and retaliation intention (‘love is blind’). → **Buffer and amplifier effect** operating at different stages of the consumer response

Note: Med. = Mediator, Mod. = Moderator, DV = Dependent Variable

Appendix 2. Literature Review – Role of Political Ideology in Controversial or Negative Brand Events

Source	Topic	Moderator(s), Mediator(s), Dependent Variable(s)	Main Findings Regarding Role of Political Ideology
Antonetti and Anesa, 2017	Reactions to corporate tax strategies	Mod.: Political identification Med.: Perceived ethicality of the firm DV: Attitude towards firm, nWOM, purchase intentions	Consumers react more negatively to aggressive tax strategies (i.e., decrease their attitude and purchase intentions, and increase nWOM) than they reward conservative tax strategies. Perceived ethicality mediates this effect, and left-leaning (e.g., liberal) consumers are more sensitive to tax strategy information as signal for firm ethicality compared to right-leaning (e.g., conservative) consumers.
Copeland and Boulianne, 2022	Meta-analysis on political consumerism	DV: Political consumerism (boycotting, buycotting)	Various factors influence political consumerism (e.g., education, age, political interest, political ideology, political trust, media use). Concerning political ideology, liberals are more likely to engage in boycott/buycott behavior (vs. conservatives).
Duman and Ozgen, 2018	Drivers of consumers' willingness to punish or reward brands associated to a political ideology	Med: Brand attitude DV: Willingness to punish/reward brand	Consumers punish brands they oppose politically, particularly when self-congruence is missing, and when brands are believed to harm societal well-being by polarizing and conservatizing the society. Moreover, religious service attendance led to stronger punishment of such brands which are associated to a political ideology.
Fernandes *et al.*, 2020	Motivations of liberals and conservatives to boycott and buycott	Med: Individualizing moral values; attitude towards boycotting DV: Boycott and buycott likelihood	Different moral concerns serve as boycott/buycott motivators: individualizing moral values for liberals (i.e., protection of harm and fairness) and binding moral values for conservatives (i.e., protection of authority, loyalty, and purity). Individualizing moral values are linked to a more positive attitude towards boycotting, explaining why liberals are more likely to engage in boycott and buycott behavior.
Garg and Saluja, 2022	Influence of brand activism (BA) issue and activism types on consumer behavior	Mod: Political ideology Med: Pride, brand value identification	Liberals (vs. conservatives) express more positive brand attitudes and higher WTP when brands engages in BA (vs. being silent). General attitude towards brand activism depends on the issue, e.g., liberals have a more positive general attitude when topic is pro-liberal (vs. pro-conservative or control). Pride and brand value identification mediate the effects of BA on consumer responses.

		DV: Brand attitudes and willingness to pay (WTP)	
Jung *et al.*, 2017	Impact of political ideology on complaining and disputing behavior	Mod: Political ideology Med: System justification motivation DV: Consumer complaint and dispute	Conservatives (vs. liberals) are less likely to report complaints and dispute complaint resolutions, due to a stronger motivation to engage in “system justification”.
Jung and Mittal, 2020	Effect of political ideology in different stages of customer journey	Mod: Political ideology DV: WOM, repurchase intentions	Political ideology has an impact on consumer decisions in different stages of customer journey in retailing. Different antecedents and correlates are reviewed, e.g., personality traits, different moral values, cognitive differences, authoritarianism or likability to engage in system justification. WOM and repurchase intentions for major US-brands are assessed depending on political ideology.
Ketron *et al.*, 2022	Influence of political ideology on reactions to brand activism (corporate political stance-taking)	DV: Patronage intentions, willingness to pay (WTP)	Liberal (vs. conservative) consumers react stronger to companies’ political stance-taking as they place higher priority on brand activism. Thus, agreeing (vs. disagreeing) to these stances lead to a wider variance of patronage intentions and WTP of liberal compared to conservative consumers.
Northey *et al.*, 2020	Consumer responses to gender and sexuality in advertisements	Mod: Political ideology Med: Disgust, attitude towards advertisement DV: Attitude towards product	For conservatives (vs. liberals), a product was less appealing when advertisements incorporated male-to-male homosexual imagery, driven by feelings of disgust. Effect was not present for female-to-female homosexuality or heterosexual couples.
Pecot *et al.* 2021	Influence of political ideology on anti-consumption	Mod.: Political ideology, market attitudes, authoritarianism DV: Anti-consumption manifestations	Anti-consumption manifestations (i.e., suspicion of marketing, intention to boycott) are stronger for politically extreme (vs. moderate) individuals and for consumers who favor stronger market regulations (vs. a liberalized market) and higher authoritarianism (i.e., paternalistic societal hierarchy).

Weber *et al.*, 2023	Consumer responses to corporate political advocacy (CPA) and CSR	Mod: Political efficacy Med: Appropriateness and Controversiality DV: Sentiment on social media, brand attitude	CPA is less appropriate and more controversial compared to CSR, leading to less favorable brand attitude. Consumers' response depends on the match or mismatch between their and firm's political orientation which is inferred from the firm's actions. Negative effects in case of mismatch were stronger than positive effects in case of match, and effects were stronger for CPA (vs. CSR). Moreover, political efficacy buffered negative effects of a mismatch.

Note. *Med. = Mediator, Mod. = Moderator, DV = Dependent Variable*

Appendix 3. Scenarios used in studies 1-3

Studies 1 and 2

During a press conference last week, in response to a question about immigrant workers, a spokesperson of [Company Name] said:

Case 1 (Pro-immigrants, adopted from Mukherjee and Althuizen, 2020): "As a company, we believe that all illegal immigrants cannot be asked to leave the country without considering how long they have already been living here. This is a matter of compassion. There should be some solution. This country has been benefited in many ways by the contributions of such illegal immigrants."

Case 2 (Contra-Immigrants, adopted from Mukherjee and Althuizen, 2020): "As a company, we believe that all illegal immigrants have to be asked to leave our country irrespective of how long they have been living here. They violated the law of the land. There should be no amnesty. The rightful citizens of our country have been badly affected for far too long by such illegal immigration."

Case 3 (Control): "As a company, we believe our success is largely caused by our people. Everyone here contributes a part - their time, motivation, joy and work - to our processes and, in turn, to the creation of something great. Our activity makes sense for all of us and also for our customers. Therefore, everyone in this company can find meaning in their daily work."

Study 3

Case 1 (Pro-Choice): According to reports, [Company Name] has been socially active and voiced their opinion regarding contemporary issues. For example, the company recently took a strong pro-choice stand in the reproductive rights debate. The company also financially contributes to the pro-abortion rights movement that actively supports Planned Parenthood.

Case 2 (Pro-Life): According to reports, [Company Name] has been socially active and voiced their opinion regarding contemporary issues. For example, the company recently took a strong pro-life stand in the reproductive rights debate. The company also financially contributes to the anti-abortion rights movement that actively opposes Planned Parenthood.

Case 3 (No-Stance): According to reports, [Company Name] has been socially active but impartial on contemporary issues. For example, the company refrained from taking a stand in the reproductive rights debate. The company also decided to not financially contribute to the sociopolitical movements, such as the pro-life or pro-choice movements.

Appendix 4. Study 3. Conditional process model with perceived marginalization as mediator and brand attitude as outcome

	Mediator model: Perceived marginalization		Outcome model: Brand attitude	
	b	*t*	*b*	*t*
X1: Disagreement (vs. agreement)	2.53	4.90 ***	-1.59	-4.44 ***
X2: No-Stance (vs. agreement)	0.38	0.69 $^{\text{n.s.}}$	-0.22	-0.60 $^{\text{n.s.}}$
M: Perceived marginalization	-	-	-0.13	-3.61 ***
W1: CBI	0.15	1.71 †	0.16	2.72 **
X1*W1 (CBI)	-0.13	-1.25 $^{\text{n.s.}}$	0.21	2.89 **
X2*W1 (CBI)	0.27	2.55 *	-0.03	-0.49 $^{\text{n.s.}}$
W2: Political ideology (PI)	0.33	5.68 ***	0.07	1.71 †
X1*W2 (PI)	-0.25	-2.65 **	0.00	0.03 $^{\text{n.s.}}$
X2*W2 (PI)	-0.30	-3.48 ***	-0.00	-0.03 $^{\text{n.s.}}$
COV: Pre-brand attitude	-0.41	-4.85 ***	0.60	10.12 ***

Conditional direct effects of X on Y (brand attitude)

X1: Disagreement (vs. agreement)	CBI	Political ideology	*b*	*t*
	weak	liberals	-1.04	-5.16 ***
	weak	conservatives	-1.03	-3.99 ***
	strong	liberals	-0.27	-0.27 $^{\text{n.s.}}$
	strong	conservatives	-0.26	-0.20 $^{\text{n.s.}}$

X2: No-stance (vs. agreement)	CBI	Political ideology	*b*	*t*
	weak	liberals	-0.32	-1.57 $^{\text{n.s.}}$
	weak	conservatives	-0.32	-1.34 $^{\text{n.s.}}$
	strong	liberals	-0.45	-1.85 †
	strong	conservatives	-0.45	-2.30 *

Conditional indirect effect(s) of X on Y (brand attitude) at values of the moderator (with 95% CI's)

X1: Brand disagreement (vs. agreement)	CBI	Political ideology	*b*	Lower	Upper
	weak	liberals	-0.21	-0.36	-0.07
	weak	conservatives	-0.07	-0.20	0.02
	strong	liberals	-0.15	-0.31	-0.03
	strong	conservatives	-0.01	-0.10	0.06

X2: No-stance (vs. brand agreement)	CBI	Political ideology	*b*	Lower	Upper
	weak	liberals	-0.06	-0.17	0.02
	weak	conservatives	0.11	-0.00	0.24
	strong	liberals	-0.18	-0.36	-0.05
	strong	conservatives	-0.02	-0.10	0.05

Note. * *p* < .05, ** *p* < .01, *** *p* < .001, † = marginally significant, n.s. = not significant.

CBI weak (-1 SD: 2.64), CBI strong (+1 SD: 6.40), liberals (-1 SD: 2.05), conservatives (+1 SD: 6.44), 95% CI's: Bootstrap 95 per cent confidence intervals for conditional indirect effect

Paper 3

AI wrote this? Effects of AI-powered content creation on employer attractiveness

Status: under review at the *European Journal of Marketing*

Co-Authors: Stefanie Wannow, Jan Freidank and Alexander Haas

Own Share: 70%

AI wrote this? Effects of AI-powered content creation on employer attractiveness

Abstract

Artificial intelligence (AI) is increasingly used to automatize content-creation and create efficiencies for firms and content creators such as news agencies. While journalism studies have investigated AI authorship and impact on their stakeholders, marketing research has largely neglected consumers' responses to AI-authored content. Using a between- and within-subjects design, this study is the first to examine the potential of AI-authored text for branding activities and to investigate readers' reactions of AI disclosure. Building on expectancy violation theory and the concept of algorithm aversion, we find that readers perceive AI-generated image texts equal to human-generated texts - as long as the author is unknown. AI disclosure elicits reader's feelings of betrayal, which negatively affect employer attractiveness perceptions. Moreover, a match or violation of a reader's authorship expectation was found to be an important factor for shaping these negative responses. We discuss implications of user reception of AI texts and present recommendations for beneficial AI integration for marketing managers.

Keywords: Artificial intelligence; Content creation; Algorithm aversion; Expectancy violation, AI transparency; AI ethics

1. Introduction

Today, many firms are confronted with a rising competitive pressure and seek ways to design processes more efficiently (Graefe & Bohlken, 2020; Wölker & Powell, 2018). Recently, artificial intelligence (AI)–empowered digital tools have been suggested to provide these process efficiencies (Davenport, Guha, Grewal, & Bressgott, 2020). AI, that is “machines that mimic human intelligence in tasks such as learning, planning, and problem-solving through higher-level, autonomous knowledge creation” (Bruyn, Viswanathan, Beh, Brock, & Wangenheim, 2020, p. 93), is getting increasingly sophisticated and can catch up with or even outperform humans today. Numerous fields and industries use and advance AI for a wide range of applications, such as medical diagnosis, autonomous driving, programmatic advertising, and trading. Among these application possibilities, the opportunity to use AI for autonomous content creation not only gains momentum in journalism, but also in the fields of marketing and human resource management (Bruyn et al., 2020; Kietzmann, Paschen, & Treen, 2018; van Esch, Stewart Black, Franklin, & Harder, 2020). Still, marketing managers hesitate to implement AI technology. They are unsure whether using AI would lead to positive reactions lifting the firm’s image as an innovative company and technological leader, or to negative reactions leading to rejections or harming the firms’ image.

Understanding consumer responses to AI-generated content becomes ever more important for marketing managers. While marketing research has largely neglected to investigate consumer responses to AI, scholars from fields as diverse as augmented decision-making, forecasts, and medical diagnoses found that people tend to distrust AI even when it delivered equal or even better service—a phenomenon labelled algorithm aversion (Dietvorst, Simmons, & Massey, 2015, 2016; Longoni, Bonezzi, & Morewedge, 2019). The risk of negative reader reactions is also closely connected to the crucial question of AI transparency. Currently, there are no legal restrictions, but calls for AI declaration as an ethical principle are becoming more frequent and intense (Jobin, Ienca, & Vayena, 2019). Likewise, several courts and organizations have called for laws to enforce declaration of AI use. For instance, the EU Commission's High-Level Expert Group on AI published AI guidelines on topics, including transparent authorship (AI HLEG, 2019). Scholars identified transparency as a prevalent quality of ethical AI (Siau & Wang, 2020). A global meta-analysis by Jobin et al. (2019) showed that more than 70 of 84 evaluated sources from business, politics, and academia included the call for AI transparency to foster trust, minimize harm, or enable human participation. This leads to the assumption that a transparent AI declaration will become the ethical standard and a legal requirement in the future.

Thus, from a theoretical and managerial perspective, the relevant questions are: Whether, why, and how do people react to transparent AI use for marketing-related applications, such as advertising or employer branding content? Furthermore, how do their authorship expectations influence their perceptions?

The present study addresses these questions by examining the effects of an AI (vs. human) authorship disclosure regarding employer branding communication on the social media platform LinkedIn. We used this study context because many firms have assigned a key priority to talent attraction and acquisition. Around 80% of CEOs are concerned about skill shortages and struggle to attract potential employees (Dögl & Holtbrügge, 2014). To overcome this challenge, they use employer branding to communicate distinctive features of their business activities, visions, and work environment to attract potential talents (Baum, Schäfer, & Kabst, 2016). In particular, social media has become a relevant platform for employer branding, as individuals increasingly rely on social media content as a source of information during job searches (Mulvey, Esen, & Coombs, 2016). Thus, content on these platforms might strongly influence a firm's image as a desirable workplace and a great employer (Kissel & Büttgen, 2015). To constantly deliver new content and save costs at the same time, the impact of AI use in this field is a fascinating question for both research and management.

To conceptualize the effect of an AI authorship disclosure on an individual's attitude toward the firm as a potential employer, we integrated two well-established theoretical concepts: expectancy violation theory (Burgoon & Hale, 1988) and algorithm aversion (Castelo, Bos, & Lehmann, 2019; Dietvorst et al., 2016). Thereby, this study integrates readers' expectations of AI vs. human authorship and a successive authorship disclosure in a two-step research design. Finally, this paper introduces feelings of betrayal as mediator between authorship disclosure and evaluations of the company's attractiveness as a potential employer.

This research makes several contributions. First, it integrates theory on AI performance from the field of journalism into the marketing field. In particular, this study is the first that uses a real company image text and a real AI-created version to evaluate brand-related consumer responses. Second, the study contributes to literature on algorithm aversion (Dietvorst et al., 2015, 2016) by showing its relevance in the field of content creation. Third, identifying the importance of expectancy violation theory and the algorithm aversion concept supports a more nuanced understanding of readers' reactions toward AI- or human-authored texts. Our two-step research design also enables the comparison between objective (i.e., without knowing the authorship type) and subjective (i.e., after authorship disclosure) AI performance in content creation, as we apply pre- and post-disclosure measurements. Finally, we shed light on the underlying affective process, showing that feelings of betrayal mediate the impact of disclosed AI authorship on employer attractiveness. Overall, we anwer the important questions of whether, why, and how people react to AI-generated communication.

From a managerial perspective, this research provides desperately needed guidance for AI implementation in image-related activities, such as social media and corporate relations communication. We show that a disclosure of AI authorship remains to be a risky strategy and that talents' responses evoked by the message author must be considered. The results of this study contribute to the current discussion on AI disclosure and ethics by measuring its effects in an important potential business application.

2. Conceptual Framework

2.1. AI Performance and Algorithm Aversion

Fueled by technological advances, such as natural language processing, neural networks, and deep learning, AI is increasingly capable of performing tasks with minimal or even without human input (Carlson, 2015; Huang & Rust, 2018). Recent studies have shown the impressive capabilities of AI. For instance, AI can detect diseases more accurately than physicians, sell as effectively as proficient salespersons, or create highly precise forecasts and personalized product recommendations (Bruyn et al., 2020; Chan-Olmsted, 2019; Longoni et al., 2019; Luo, Tong, Fang, & Qu, 2019). Therefore, AI applications are considered the governing technology of the digital age (Chan-Olmsted, 2019).

Given the equal or even superior performance of many AI-driven applications in forecasting, providing advice, or customer support, consumers could similarly rely on AI or even more than on humans (Longoni et al., 2019). However, scholars found that people do not. AI research has therefore defined this the phenomenon as *algorithm aversion*, asserting that individuals generally prefer humans over AI (Castelo et al., 2019; Dietvorst et al., 2015, 2016). Many customers were found to have reservations toward AI services; they feel uncomfortable being served by AI, which may attenuate their attitudes and subsequent behaviors toward the firm (Davenport et al., 2020; Longoni et al., 2019). Luo et al. (2019) suggests that this negative effect resulting from the disclosure of AI seems to be "driven by a subjective human perception against machines" (p. 2) despite an objectively high AI performance.

Scholars suggested that several factors are driving algorithm aversion, such as the belief that AI is unable to consider qualitative data or individual goals, lack of trust in AI, or the desire for perfect results (Dietvorst et al., 2016; Grove & Meehl, 1996; van der Kaa & Kramer, 2014). Furthermore, despite an objectively high AI performance, individuals perceived an AI—

compared to a human—as less empathetic and less competent (Luo et al., 2019). Some also fear that AI is threatening jobs or might take over too much control in their lives (Huang & Rust, 2018).

2.2. Readers’ Perceptions of AI-Created Content

As one of the first businesses, news agencies have started to integrate AI in their content creation processes to increase productivity levels and save costs compared to human journalism (Bucher, 2017; Carlson, 2015). Scholars expect that the industry for computer-written news will expand quickly, and automated journalism is even likely to disrupt news creation in the future (Wu, 2019). Firms from various industries are following suit. They seek to leverage value from AI-enabled automation in advertising, sales, marketing, or recruitment (Kietzmann et al., 2018; van Esch et al., 2020). Yet, related studies remain limited and can mainly be found in journalism (see Table 1 for an overview).

[Insert Table 1 about here]

Most studies focus on perceptual or cognitive outcomes of AI-created (vs. human-created) content, such as content quality or credibility (Clerwall, 2014; Graefe, Haim, Haarmann, & Brosius, 2016; Waddell, 2018; Wölker & Powell, 2018; Wu, 2019). As an exception, Wölker and Powell (2018) also examine a behavioral intention. They assess whether the disclosure of an AI or human author influences readers’ news selection, and found no significant impact of AI on news selection or avoidance.

Effects of AI Authorship disclosure

Most studies disclose the content authorship directly with the content and measure the effects of authorship disclosure. Effects of AI authorship disclosure were found to be mixed. Some scholars find negative effects (e.g., lower author credibility; Waddell, 2018) or no differences in readers' perceptions (Wölker & Powell, 2018). Other research assert also positive effects, for instance, respondents were positively surprised about AI's high level of content quality because they expected a lower performance level and accuracy (Graefe et al., 2016). In support of this, journalism practice reports acknowledge the increasing capabilities of recent AI tools for text creation (Hansen, Roca-Sales, Keegan, & King, 2017).

Without author disclosure, a study using short texts on sports results show that people were unable to detect the source, supporting the good quality of AI texts (Clerwall, 2014). Comparing disclosure of authorship (yes vs. no) and source information (e.g., *AI from Forbes*), Wu (2019) revealed that a message without a disclosed source and author was rated more credible than one with a mentioned source and author. As to the mechanism, only two studies evaluate mediating effects (Waddell, 2018; Wölker & Powell, 2018), and only Waddell (2018) found a significant mediation effect, i.e., the influence of AI authorship on content credibility is mediated by lower perceptions of AI (vs. human) anthropomorphism. And related to readers' perceptions of automated vs. human-written news, the recent meta-analysis from Graefe and Bohlken (2020) summarized 12 studies and found strong advantages for human-written texts (vs. AI) regarding readability and minor advantages for quality perceptions, but no differential effects for credibility. Thus, although the results remain somewhat inconclusive, there are cues indicating algorithm aversion in content-related research.

Expectancy violations

Next to aversive perceptions of AI, individual expectations about the "nature" of one's interaction partner and its (dis)confirmation might affect user reactions. Several theories in communications, marketing, and psychology integrated the principle of the confirmation or disconfirmation of expectations (Burgoon & Hale, 1988; Oliver, 1980). Among them, the *expectancy violation theory* (Burgoon & Hale, 1988) proposes that an individual's behavior that deviates from one’s currently held expectations and social norms will trigger (1) emotions and (2) attempts to explain this behavior. Although this theory is rooted in interpersonal communication, scholars applied it across several domains, such as information technology or corporate messages (Burgoon et al., 2016; Rim, Park, & Song, 2020).

Expectations help to make one’s life easier and manageable. They form cognitive structures to understand and interpret others’ behaviors and influence mental information processing. These expectations are found to influence consumer evaluations or interpersonal communications. A person builds expectancies based on another actor’s characteristics, the relationship to one another, and the context surrounding the interaction (Burgoon, 2009).

In general, expectations could be confirmed or violated (Burgoon & Hale, 1988). In turn, a violation of prior held expectations can result in positive or negative reactions. This depends on the (subjectively rated) direction of the enacted behavior related to the expected behavior (Burgoon et al., 2016). That means, a violation can be viewed favorably (i.e., positive violation) or unfavorably (i.e., negative violation). Furthermore, when expectancies are violated, people’s attention is shifted from the content to the message sender and the violation’s salience increases (Burgoon & Hale, 1988). That means people focus their attention more toward evaluating the violation and try to interpret possible reasons behind the discrepancy.

Regarding content authorship, readers build certain expectations of the author as the source of a text. With the advent of AI authorship and first examples in the media, it gets more common today for readers to assume also an AI author for a specific text. In turn, two forms of expectancy violation could occur. Either a person expected a human author and received an AI generated text, or vice versa. These expectancy violations can lead to different positive or negative emotional reactions, and subsequent effects towards the firm as message sender.

2.3. Creating Social Media Content for Employer Branding

At the interface between marketing and human resource management, one promising application of AI is content creation for employer branding campaigns. Employer branding, defined as a firm's strategic efforts to transfer a positive image as a good place to work (Backhaus & Tikoo, 2004), has become a source of competitive advantages for firms (Dabirian, Kietzmann, & Diba, 2017). A major target of employer branding is to create an attractive company image, which transfers a unique employer "value proposition" (Backhaus & Tikoo, 2004, p. 502) to attract high-potential future employees. Texts such as job advertisements or social media profiles serve as employer branding instruments to increase employer attractiveness perceptions (Kissel & Büttgen, 2015). Employer attractiveness refers to an "individual's affective and attitudinal thoughts about particular companies as potential places for employment" (Highhouse, Lievens, & Sinar, 2003, p. 989).

Firms can transfer employer branding messages and content through different channels. The proliferation of new digital platforms reshaped how firms interact with job seekers and where potential employees are searching for information. Social media platforms, including Facebook, Instagram, and LinkedIn, as most notable ones, have seen an enormous rise in usage and attractiveness for both firms and prospective talents. In particular, LinkedIn, the major international business network platform, is used by 94% of the world's 500 fastest-growing companies for talent acquisition and is the most effective website for recruitment for three of four human resource professionals (Mulvey et al., 2016). Therefore, providing up-to-date and relevant information in an efficient and timely manner (e.g., by using AI-created content) becomes a crucial success factor for employer branding campaigns. Due to its high relevance and the lack of research in this particular area of application, we used employer branding as the context of our study and conceptualized employer attractiveness as the main dependent variable.

2.4. Hypotheses Development

To understand the effects of type of authorship (AI vs. human) on employer attractiveness, we first focus on its effects without author disclosure. Today, AI tools are increasingly able to produce texts with compelling, entertaining, and informative content (Bruyn et al., 2020; Graefe & Bohlken, 2020). Indeed, initial studies provide evidence that AI is already so advanced that individuals could not differentiate between automated and human-generated content or rate AI texts even better in some domains (Graefe & Bohlken, 2020; Wu, 2019). Thus, regarding an employer branding message from an AI vs. human author, we expect readers to perceive the firm's employer attractiveness equally positive.

Authorship Expectations

However, current studies have not integrated readers' expectations about the author type. As AI-written content is on the rise, some readers might even expect an algorithm instead of a human author. In line with AI aversion, a mere expectation of an AI (vs. a human) author should negatively influence employer attractiveness evaluations. More specifically, while we suggested that objective AI vs. human authorship does not influence readers' perceptions in the pre-disclosure stage, we predict that expected authorship will make a difference. Thus:

H1. *AI (vs. human) authorship expectation has a negative effect on employer attractiveness in the pre-disclosure stage.*

Effects of Authorship Disclosure

In case of authorship disclosure, the type of author (AI vs. human) reflects an additional piece of information for an individual that will be integrated in the subjective evaluation (Clerwall, 2014; Wölker & Powell, 2018). This means that readers add this information to refine their assessment of the text and about the sender of the message. Drawing on the algorithm aversion construct, we expect a negative effect of AI (vs. human) authorship disclosure. In particular, we suggest that AI authorship disclosure reduces employer attractiveness. Scholars showed that people have negative associations with AI, and tend to distrust AI's accuracy and credibility (Luo et al., 2019; Waddell, 2018). Journalism studies also show that readers tend to rate an AI author as less valuable or credible (Graefe et al., 2016; Graefe & Bohlken, 2020; Waddell, 2018). Consequently, we expect disclosure of an AI author to lead to a less favorable employer attractiveness (Highhouse et al., 2003). We hypothesize:

H2. *AI (vs. human) authorship disclosure has a negative effect on employer attractiveness.*

The Role of Readers' Feelings of Betrayal

The role of emotions in marketing and communications has been investigated extensively (Bagozzi, Gopinath, & Nyer, 1999; Poels & Dewitte, 2019). Following Bagozzi et al. (1999), emotions are an important mediator between marketing and communication stimuli and individuals' responses. When being confronted with employer branding messages and image texts, individuals do not only cognitively process the presented information and its source but also respond emotionally (Baum et al., 2016; Rampl, Opitz, Welpe, & Kenning, 2016). Furthermore, following related literature, AI as content author is asserted to trigger mainly negative emotions (Waddell, 2018).

Following related AI literature and the algorithm aversion construct, individuals are mainly supposed to express negative emotions when reading AI-created content. In particular, scholars show that individuals have mainly negative associations with AI (Huang & Rust, 2018; Luo et al., 2019). For instance, they feel less valued as customers because the process automation might signal less effort and resource allocation to their favor (Huang & Rust, 2018; Waddell, 2018). And more generally, Castelo et al. (2019) showed that individuals felt more negative and less in control when an AI takes over a certain task instead of a human. Among these negative emotions, a particularly strong one is the feeling of *betrayal*, which describes a “customer’s belief that a firm has intentionally violated what is normative in the context of their relationship” (Grégoire & Fisher, 2008, p. 250). Previous research has shown detrimental effects of individuals’ feelings of betrayal, for instance, on customers revenge intentions, negative word-of-mouth, or churn (Grégoire & Fisher, 2008; Riquelme, Román, Cuestas, & Iacobucci, 2019). In the context of employer branding, a job reflects an intense relationship with a firm (Baum et al., 2016). Therefore, job seekers are likely to pay more attention toward the credibility of the content provided, and feelings of betrayal might create a quite severe impact for the firm. Particularly regarding a text that describes the identity or “character” of a firm and is rather emotional compared to factual texts, such as a weather forecast, people expect a more personal touch. Furthermore, assessing a firm as a potential employer might represent a quite important or even risk-related topic, as this decision affects peoples’ lives. Therefore, readers might see it as normative that humans handle this topic, and the violation of this kind of norm should trigger betrayal. Especially betrayal seems to be a likely response, because intended norm violations are one of its defining characteristics (Burgoon, 2009). This negative feeling of betrayal should, in turn, influence text and firm evaluations.We postulate:

> **H3:** *The effect of authorship disclosure on employer attractiveness is mediated by reader’s feeling of betrayal.*

The Influence of Violated or Matched Expectations

When a content author is disclosed, a reader compares this new piece of information with pre-held expectations and a mental processing is started (Burgoon & Hale, 1988). In particular, readers’ expectations about a text’s author could either be matched, i.e., they correspond with the assumed author; or violated, i.e., they do not correspond with their assumption (Burgoon, 2009).

Regarding content creation, four distinct match or violation scenarios are possible: As mentioned, readers might traditionally expect a human author – as this reflects the current norm and standard case (Waddell, 2018). This expectation could be matched (i.e., disclosed human author) or violated (i.e., disclosed AI author). In contrast, readers might also expect an AI author, as news start to report about AI use for content creation in news agencies and firms (Chandler, 2020; Marr, 2019). Particularly in digital environments such as social media platforms or online news websites, AI is increasingly employed and especially tech-savvy individuals could assume that AI is already in place (Marr, 2019). Again, this expectation could be matched (disclosed AI author) or violated (disclosed human author).

Following the expectancy violation theory, a negative (positive) expectancy violation triggers mental processing and leads to negative (positive) reactions (Burgoon & Hale, 1988). Drawing on the algorithm aversion concept, when people compare AI vs. human authorship, an AI author represents a less desired and less trustworthy option (Luo et al., 2019). Thus, when people expect a human author, but receive the less desired option "AI" – this disclosure should result in negative emotional reactions. More precisely, this violation should lead to stronger feelings of *betrayal* (than not violated) (Waddell, 2018). In contrast, when readers assume an AI author and receive a human author, this represents a (positive) violation with the more desired option of human authorship but at the same time the common author type today (Graefe & Bohlken, 2020; Waddell, 2018). Thus, we anticipate less feelings of betrayal. We hypothesize:

> **H4:** *Expectancy violations moderate the effect of AI disclosure on readers' feelings of betrayal. More precisely, an unexpected AI (human) disclosure leads to more (less) feelings of betrayal than an expected AI (human) disclosure).*

Next to the influence emotional responses like betrayal, previous AI-studies support the notion that an AI disclosure leads to further negative attitudinal and behavioral responses (Luo et al., 2019; Waddell, 2018). People were found to devalue AI's work due to various reasons, including the mentioned lower level of trustworthiness, or feelings of missing 'skills' or empathy, sometimes despite a high objective performance (Luo et al., 2019). Likewise, related to content, disclosing AI (vs. human) as author lead to lower quality and credibility perceptions (Graefe & Bohlken, 2020). These negative content-related evaluations could also be expected to affect source-related variables, such as the firm's image as attractive employer (Baum et al., 2016; Kissel & Büttgen, 2015; Wu, 2019). Therefore, we expect:

> **H5:** *Expectancy violations moderate the effect of AI disclosure on employer attractiveness. More precisely, an unexpected AI (human) disclosure leads to more negative (positive) employer attractiveness ratings than an expected AI (human) disclosure.*

Figure 1 presents our resulting conceptual model.

[Insert Figure 1 about here]

4. Empirical Study

Text Generation and Study Context

For our text generation with AI, we chose Open AI's GPT-2 content generation model as a state-of-the-art AI application (Köbis & Mossink, 2021; Solaiman, Clark, & Brundlage, 2019). This model is trained on a diverse dataset throughout several domains and therefore capable to produce more conclusive and convincing texts than other AI models. Moreover, the model imitates and adapts to content and text style (Radford et al., 2019).

For a scenario, we selected a LinkedIn profile from a large German industrial firm rated as one of the top 10 German employers by LinkedIn (Lobosco, 2019). We chose this context, because social media is a popular platform for firms' content marketing activities. LinkedIn is the world's largest social media platform for career opportunities (Mulvey et al., 2016), making it highly relevant for employer branding activities. All brand names were deleted to exclude confounding effects. Both texts (Appendix 1) contained almost the same number of words (241 vs. 239 words).

Research Design and Participants

To analyze the influence of artificial- or human-intelligence-authored image texts on firms' attractiveness as employers, we created an online experiment and combined a between-subjects and a within-subjects design. Specifically, we used the AI- and human-generated texts, and measured employer attractiveness before and after author disclosure.

We recruited participants via Amazon MTurk. To ensure data quality, participants had to achieve an approval ratio of at least 95% in former tasks. Based on realism- and attention-checks (Wagner, Hennig-Thurau, & Rudolph, 2009), fourteen participants were excluded from the data analysis, resulting in a usable sample of 445 participants (M_{Age}=36 years, 41% female). The majority (71%) was active on LinkedIn, and social media seemed to play an important role for most of them. For instance, social media was part of their everyday activities (59% rated 5 or above on a 7-point scale), and they followed companies and their brands on social media (62% rated 5 or above).

Participants were randomly assigned to one of two scenarios: the original image text from a human author (n_{Human}=220) and the AI-generated version (n_{AI}=225). We asked them to imagine looking for a new job. First, we showed the text to the participants without disclosing the author. Second, we asked them to rate the text and the company. Subsequently, the respondents chose their authorship expectations. To disguise the real target of the study, we offered three, instead of two, possible author categories, namely, native writer, non-native writer related to human authorship, and AI/machine as AI authorship. Third, after author disclosure participants responded to the mediator and attractiveness scales of our model. We controlled for respondents' socio-demographics (i.e., age, gender, level of education), social media usage, and possession of LinkedIn accounts. In both pre- and post-disclosure conditions, all mean values of these control variables were not significantly different between the human and AI-author groups (each $p > 0.1$).

Measures

Employer attractiveness was assessed based on five items (Highhouse et al., 2003). A sample item is: "For me, this company would be a good place to work." The response format was a 7-point Likert scale ranging from 1 (fully disagree) to 7 (fully agree). A reverse-labeled item had to be dropped due to insufficient factor loadings. Betrayal was assessed with two items: "I feel fooled" and "I feel tricked" (Darke, Ashworth, & Ritchie, 2008). Expectancy match or violation was calculated by comparing the objective authorship type and the expected author. Construct means and reliability measures can be found in Appendices 2 and 3.

Results

Pre-disclosure Evaluation

We started with an analysis of pre-disclosure effects. As expected, the respondents from both groups described the texts equally as highly realistic (M_{Human}: 5.73, M_{AI}: 5.89 on a 7-point scale, $t(445) = -1.36$, $p > 0.05$). In general, the employer attractiveness of both texts was rated as positive (M_{Human}: 5.44, M_{AI}: 5.48, $t(445) = -0.40$, $p > 0.05$). Regardless of the objective authorship, most of the respondents thought that the text was written by a human writer and chose native writer (62.5%) or non-native writer (11.9%), while around a quarter suspected AI authorship (25.6%). The split between correct vs. wrong guesses of authorship was almost perfect halved (48.5% vs. 51.5%).

To test H1, we ran a two-way ANOVA to test the influence of objective and expected authorship on Employer Attractiveness (EA) prior to authorship disclosure. Objective AI vs. human authorship did not have any significant effects on the employer attractiveness rating (M_{AI}: 5.48; M_{Human}: 5.44, $t(443) = -0.40$, $p > 0.10$). In contrast, expected AI authorship affected employer attractiveness negatively ($M_{ExpectedAI}$: 5.06; $M_{ExpectedHuman}$: 5.60, $t(443) = 3.59$, $p < 0.001$). In sum, AI-written texts perform objectively equal, but a reader's expectation of AI authorship leads to negative employer attractiveness ratings, supporting H1.

Post-disclosure Evaluation

To test the authorship disclosure effect (H2), we ran a mixed ANOVA with the dependent variable employer attractiveness as a repeated measure and authorship (AI vs. human) as a between-subjects factor. The results revealed a significant main effect of disclosure on employer attractiveness ($F(1,443) = 33.92$, $p < 0.001$). Furthermore, we found a significant interaction effect of authorship and disclosure ($F(1,443) = 34.81$, $p < 0.001$), indicating that the effects of disclosure depend on whether the author was identified as AI or human. Using subsequent paired sample t-tests with a bootstrapping sample of 5,000, we then investigated the differences in the AI and human author group separately. In the AI condition, participants rated employer attractiveness significantly worse after the disclosure ($p < 0.001$). However, in the human text group, no significant changes in employer attractiveness could be observed ($p > 0.05$). All in all, these results confirm H2 (see Figure 2).

[Insert Figure 2 about here]

To measure the mediating effect of *betrayal* on the effect of author disclosure and employer attractiveness (H3) and the interactive effects of authorship and expectancy violations on betrayal (H4) and employer attractiveness (H5), we used a moderated mediation analysis through PROCESS model 8 (Hayes, 2018) with 5,000 bootstrapped samples and 95% confidence intervals (CIs). In the model, authorship (0 = human, 1 = AI) was the independent variable, expectancy violation (0 = match, 1 = violation) was the moderator, respondents' feelings of betrayal was the mediator, employer attractiveness was the outcome variable. To measure Employer Attractiveness, we used the difference in employer attractiveness ratings between pre- and post-disclosures.

As expected in the mediation hypothesis (H3), AI authorship significantly increased consumers' feelings of *betrayal* ($b = 0.99$, $p < 0.005$). In turn, these negative feelings decreased employer attractiveness ($b = -0.07$, $p < 0.001$). Thus, H3 is supported – an AI disclosure tends to elicit negative feelings of betrayal.

Testing H4 (i.e., the effects on betrayal), the interactive effect of authorship x expectancy violation on feelings of *betrayal* was insignificant ($p = .61$). Notably, expectancy violation had a considerable direct effect on *betrayal* ($b = 0.89$, $p < 0.005$). Thus, for both possible expectancy violation conditions, people seemed to feel more betrayed than when they received the author they expected. H4 could not be supported. Accordingly, the indirect effect of AI authorship on employer attractiveness through feelings of *betrayal* was significant in the case of non-violation ($b = -0.07$, 95% CI = [-0.15, -0.02]) and in the case of expectancy violation ($b = -0.05$, 95% CI = [-0.12, -0.01]), as the CIs did not include zeros. Subsequently, the index of moderated mediation was insignificant ($b = 0.02$, 95% CI = [-0.05, 0.09]). This supports the results of H4 again and shows that people feel betrayed irrespective whether they expected AI as author or not.

Regarding the moderation analysis with *employer attractiveness* as dependent variable (H5), the interactive effect of authorship x expectancy violation on employer attractiveness was significantly negative ($b = -0.53$, $p < 0.01$). Interestingly, the main effect of expectancy violation was insignificant ($p = 0.14$), showing that the effect of expectancy violation relied upon the authorship. The (negative) direct effect of AI on employer attractiveness was significant in the case of "violated expectations" ($b = -0.67$, $p < 0.001$, [-0.94, -0.39]). In contrast, this effect became insignificant when expectations were matched ($p = 0.34$) indicating comparable effects on EA ratings for both the AI and human text as long as the specific authorship type was expected.

As H5 hypothesized a positive (negative) effect of an expectancy violation on employer attractiveness in case of a human (AI) author, we needed to dive deeper into the effects of each author. In particular, we splitted the dataset into (1) human and (2) AI authorship and used a repeated measures ANOVA for each subset. Specifically, we compared the pre-and post-disclosure ratings of employer attractiveness, separated by expectancy violation. Results for the (1) human text showed no effects of (pre-vs. post-) disclosure ($p = 0.47$) or disclosure * expectancy violation ($p = 0.15$) on employer attractiveness. In contrast, results for the (2) AI text showed significant negative effects of disclosure ($p < 0.001$) and disclosure * expectancy violation ($p < 0.05$). In sum, H5 is partly supported. When people received an unexpected AI author (i.e., the asserted negative violation), this condition lead to a stronger decrease of employer attractiveness ratings than an expected AI author (i.e., $\Delta_{\text{unexpected AI}} = -0.62$ vs. $\Delta_{\text{expected AI}} = -0.25$). However, an unexpected human author (i.e., the asserted positive violation) did not lead to a (significantly) higher rating of employer attractiveness compared to an expected human author (i.e., $\Delta_{\text{unexpected human}} = +0.11$ vs. $\Delta_{\text{expected human}} = -0.04$).

[Insert Figure 3 about here]

Robustness Check

This study employed a real social media text and a real AI-generated text to enhance external validity. However, to rule out possible confounding effects stemming from the different texts itself, we conducted a robustness check (see Table 2). To do this, after the disclosure, the readers were made to believe that the AI-generated text they had just read was written by a human author. Based on another sample of 277 respondents from Amazon MTurk, we replicated the data analysis and compared this third group, labeled “fake human”, with the AI and human text groups from the main study above.

To measure effects of this new created group, we used an MANOVA with the three groups as factor and our relevant variables as dependent variables (i.e., feelings of betrayal and employer attractiveness ratings pre-and post-disclosure). Table 2 illustrates the mean values, standard deviations and results of the multiple comparisons test.

As expected, before the authorship disclosure, all three groups perceived employer attractiveness as similar, as no significant differences could be detected (each $p > 0.1$).

After the authorship disclosure, the same pattern as in the main analysis could be observed. Both the "fake human" and real human groups expressed similar feelings of betrayal ($p > 0.1$). Likewise, respondents expressed an equal employer attractiveness ($p > 0.1$). However, compared to the AI text group with AI as labeled author, both groups expressed significantly lower feelings of *betrayal* ($p < 0.001$), and evaluated also the *employer attractiveness* significantly lower ($p < 0.001$).

These results confirm that our main results are robust and that the labeling of AI or human authorship—and not the text differences—drives readers' changes in attitude.

[Insert Table 2 about here]

5. Discussion

Firms increasingly face talent attraction issues—some even call this the "war for talent" (Charbonnier-Voirin, Poujol, & Vignolles, 2017). At the same time, they are urged to streamline their processes to increase efficiencies and stay competitive in their respective markets (Baum et al., 2016). In various domains, AI applications have become an opportunity for businesses to complement or take over tasks from humans (Huang & Rust, 2018; Longoni et al., 2019) and, thus, save costs while the output remains stable. Therefore, this study investigated the objective performance and subjective perceptions of AI-created texts for talent attraction and employer branding.

Theoretical contributions

Without an explicitly labeled AI authorship, the readers evaluated an AI-written text comparable to a human-written text. This supports related studies and indicates the convincing capabilities of AI generated texts (Graefe et al., 2016; Hansen et al., 2017; Solaiman et al., 2019). In contrast to most of these studies, we used a real text from social media and compared it with a real AI-written version. Therefore, this study's first contribution is that it adds another type of text to the growing body of AI text performance literature because a longer, image-relevant, and more emotional-driven text was chosen. Several sources acknowledge that AI is suitable for rather short content that relies on factual components and structured data, such as numbers, results, or statistics (Carlson, 2015; Davenport et al., 2020; Thurman, Dörr, & Kunert, 2017). For longer and more detailed texts, however, AI needs to learn more contextual understanding. Although researchers have noted that AI seems to struggle with context awareness, this study indicates that contemporary AI applications can overcome this issue and build conclusive and context-aware content.

Second, this study adds to the literature on algorithm aversion (Castelo et al., 2019; Dietvorst et al., 2015, 2016) in several ways. To the best of our knowledge, this phenomenon has not been examined before both in the area of content creation as a promising field for AI application nor in the context of branding. Supporting the major trend of former studies, our case showed that AI disclosure (vs. non-disclosed) results in more negative brand-related evaluations, such as employer attractiveness, which are supposed to reduce application intentions of potential talents (Baum et al., 2016). Furthermore, current AI studies observing algorithm aversion (Castelo et al., 2019; Dietvorst et al., 2015, 2016) neglected to analyze individual expectations; that is, as in this case, whether a human or an AI author is expected.

By integrating this aspect, we illustrate that expectations indeed play an important role. Independent from real authorship, the readers rated the text and the firm more negatively in the pre-disclosure stage if they expected an AI author. This effect was also observed consistently in the post-disclosure stage, supporting a stable negative bias against AI authors. That means that people's expectations should be considered in future investigations of individuals' reactions toward AI.

Third, this study applies the expectancy violation theory in the digital- or AI-domain. Studies observing expectancies in the field of digital interactions or AI remain scarce. As an exemption, Burgoon et al. (2016) determined whether embodied digital agents deviated from consumers' expectations and measured the effects of these positive or negative violations. Our study sheds light on (1) the effects of users' author expectations and (2) the effects of a match vs. a violation of these expectations. As mentioned, (1) expectations influence even text evaluations without disclosed authorship, and a disclosure creates asymmetric effects. In the case of an (2) unexpected (revealed) AI author, people reacted quite negatively, whereas an unexpected human author did not change their employer attractiveness ratings.

Fourth, we uncovered the underlying effects of readers' changing perceptions. The study showed that readers' feelings impact their employer attractiveness ratings. Specifically, the negative feeling of betrayal was triggered by an AI authorship. It seems that people might feel less appreciated by the company through their encounter with AI instead of a human.

Furthermore, the study shows that author and expectancy violations interact with each other. More precisely, reading a text from a (disclosed) unexpected AI author leads to a strong devaluation of employer attractiveness and has a more negative effect than when people expected the AI authorship before. According to the expectancy violation theory, the opposite case (i.e., an unexpected human author) should create positive responses as people receive a generally more favored option than expected (i.e., human instead of AI). Although this effect was found as a slight trend in the data (see Figure 3 as well), the effect was not big enough to be significant. We assume that human authorship still represents the vast standard, and should therefore be considered as more probable and common. In turn, people might favor this option but find it not extraordinary and therefore do not increase their employer attractiveness rating.

In contrast, people felt betrayed irrespective whether the disclosed author matched their expectation or not. The direct path of AI (vs. human) authorship on betrayal was found to be a stronger predictor than possible violations (or matches). In other words, algorithm aversion has an impressive negative effect and overlayed the (generally positive) awareness of a matched author expectation.

In total, readers of a transparent AI text felt betrayed and subsequently devalued employer attractiveness ratings compared to their initial rating without authorship information. This indicates that AI aversion, at least in this context, is mainly an affective phenomenon and not based on rational considerations or objective AI performance.

Managerial implications

This study has implications for marketing and human resource managers in a digital environment. First, we show that the quality of AI content creation can now compete with traditional human content creation. AI is also technically capable of creating longer and image-relevant texts. As both texts create the same positive company image as attractive employer, firms might start to publish AI-created texts within their social media profiles to enhance an efficient content creation process. Likewise, the high objective quality of AI-created content offers firms the opportunity to create more frequent or highly personalized content without using immense human labor.

Second, this study reveals that disclosing an AI authorship is an ethical dilemma. On the one hand, identifying AI authorship reflects a firm's transparent and ethical behavior (Jobin et al., 2019; Montal & Reich, 2017). On the other hand, this disclosure leads to aversive emotions and negatively affects firm image and talent attraction. These insights provide managers with a more nuanced understanding about the consequences and underlying emotions of AI use in content marketing. These results help them to be prepared to make strategic decisions when laws make AI transparency compulsory or when firms request their managers to abide by their own transparency guidelines.

As managers know now about the mainly negative response of AI disclosure, they might engage in lobbying efforts to hinder or at least to delay laws which enforce transparency of AI use. As further dark side, some managers might even take advantage of the readers' inability to distinguish an AI from a human text. They could (mis-)use the positive associations of a human author by wrongly declaring an AI authorship as human.

As ethical alternative, firms could be "offensively transparent" regarding AI use, and create marketing campaigns to emphasize their innovativeness (i.e., using AI) and highly ethical behavior, (i.e., declaring AI use). Future studies might evaluate, whether and how these campaigns could mitigate the negative consequences of AI disclosure.

This study is even relevant for managers who do not use AI for their employer branding or advertising campaigns. People devalue a firm as attractive employer (already) when they expect that content was created by a machine. Therefore, when using human writers, managers should transparently show that a human author – instead of AI - has written the content. This should prevent the firm from the outlined negative consequences, and might even create a sense of appreciation for their own staff and future talents. Furthermore, the considerable direct impact of expectancy violations on employer attractiveness showed that firms are generally better off when they declare any kind of authorship transparently.

Taking a broader perspective, other stakeholders such as policymakers and nongovernmental organizations, might also benefit from these insights. For instance, Amnesty International, one of the authors of the Toronto Declaration, called for the absolute transparency of AI applications to prevent "discriminatory or other rights-harming impacts" (Amnesty International & Access Now, 2018, p. 9). Our findings show that AI is no longer distinguishable from a human author and can create sophisticated emotional texts. Furthermore, as soon as people detect their inability to distinguish AI authored content, they might develop a general distrust in organizations and institutions as the AI technology becomes more widespread and 'all information' might be generated from AI. Both aspects support the mentioned calls for transparency of these stakeholders, as transparency would protect people from making wrong assumptions and would preserve trust in organizations and firms.

Limitations and future research

This research offers valuable insights, but it also has limitations. First, our study is based on one particular firm image text. Future research may broaden our results' generalizability by replicating the study using company texts from other industries, countries, and media platforms.

Second, although our experimental setting used a real LinkedIn text and a real AI-created text to stay very close to reality, respondents were not in their LinkedIn accounts and not in a job application situation. By creating a study in a LinkedIn profile of a firm and integrating behavioral variables such as click rates or applications, a future study would consider real reactions to AI and add further validity to our findings.

Third, our manipulation of AI authorship considered only the two extreme options of human vs. AI authors. However, mixed approaches are also possible, and several researchers see a high probability in using AI to assist human managers in content creation (Graefe et al., 2016; Hansen et al., 2017; van Dalen, 2012). Future studies could investigate the optimal mix of AI and human content co-creation.

Fourth, this study presented the AI authorship without giving any additional information or promoting AI. Future studies could analyze whether marketing messages could actively influence readers' expectations and mitigate negative responses by providing their customers with an explanation about their approach and reasons behind their AI implementation. For instance, an introduction to a marketing message might start by stating that the company, being innovative, uses AI to explore new tools and ways of communication. A transparent AI use could give firms also a positive image as pioneers of ethical digital behavior (Siau & Wang, 2020). As a full declaration of AI authorship is likely to be mandated in the future, we suggest that future studies investigate concepts to reduce the negative emotional effects of AI authorship or even to find ways to create positive consumer reactions.

Acknowledgements

We gratefully thank Sarah Simon from Factor Eleven GmbH for her valuable ideas and support in creating the artificial intelligence text.

6. References

AI HLEG (2019). On Artificial Intelligence - A European approach to excellence and trust: The European Commission. Retrieved from https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Amnesty International, & Access Now (2018). The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems. Retrieved from https://www.torontodeclaration.org/declaration-text/english/

Backhaus, K., & Tikoo, S. (2004). Conceptualizing and researching employer branding. *Career Development International*, *9*(5), 501–517. https://doi.org/10.1108/13620430410550754

Bagozzi, R. P., Gopinath, M., & Nyer, P. U. (1999). The Role of Emotions in Marketing. *Journal of the Academy of Marketing Science*, *27*(2), 184–206. https://doi.org/10.1177/0092070399272005

Baum, M., Schäfer, M., & Kabst, R. (2016). Modeling the Impact of Advertisement-Image Congruity on Applicant Attraction. *Human Resource Management*, *55*(1), 7–24. https://doi.org/10.1002/hrm.21652

Bruyn, A. de, Viswanathan, V., Beh, Y. S., Brock, J. K.-U., & Wangenheim, F. von (2020). Artificial Intelligence and Marketing: Pitfalls and Opportunities. *Journal of Interactive Marketing*, *51*, 91–105. https://doi.org/10.1016/j.intmar.2020.04.007

Bucher, T. (2017). ‘Machines don’t have instincts’: Articulating the computational in journalism. *New Media & Society*, *19*(6), 918–933. https://doi.org/10.1177/1461444815624182

Burgoon, J. K. (2009). Expectancy Violations Theory. In C. R. Berger & M. E. Roloff (Eds.), *The International Encyclopedia on Interpersonal Communication*. New York, NY, USA: Wiley and Blackwell.

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E., & Giboney, J. S. (2016). Application of Expectancy Violations Theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies*, *91*, 24–36. https://doi.org/10.1016/j.ijhcs.2016.02.002

Burgoon, J. K., & Hale, J. L. (1988). Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communication Monographs*, *55*(1), 58–79. https://doi.org/10.1080/03637758809376158

Carlson, M. (2015). The Robotic Reporter. *Digital Journalism*, *3*(3), 416–431. https://doi.org/10.1080/21670811.2014.976412

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Chandler, S. (2020). Reuters Uses AI To Prototype First Ever Automated Video Reports. Retrieved from https://www.forbes.com/sites/simonchandler/2020/02/07/reuters-uses-ai-to-prototype-first-ever-automated-video-reports/

Chan-Olmsted, S. M. (2019). A Review of Artificial Intelligence Adoptions in the Media Industry. *International Journal on Media Management*, *21*(3-4), 193–215. https://doi.org/10.1080/14241277.2019.1695619

Charbonnier-Voirin, A., Poujol, J. F., & Vignolles, A. (2017). From value congruence to employer brand: Impact on organizational identification and word of mouth. *Canadian*

Journal of Administrative Sciences / Revue Canadienne Des Sciences De L'administration, *34*(4), 429–437. https://doi.org/10.1002/cjas.1379

Clerwall, C. (2014). Enter the Robot Journalist. *Journalism Practice*, *8*(5), 519–531. https://doi.org/10.1080/17512786.2014.883116

Dabirian, A., Kietzmann, J., & Diba, H. (2017). A great place to work!? Understanding crowdsourced employer branding. *Business Horizons*, *60*(2), 197–205. https://doi.org/10.1016/j.bushor.2016.11.005

Darke, P. R., Ashworth, L., & Ritchie, R. J. (2008). Damage from Corrective Advertising: Causes and Cures. *Journal of Marketing*, *72*(6), 81–97.

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, *48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology. General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dögl, C., & Holtbrügge, D. (2014). Corporate environmental responsibility, employer reputation and employee commitment: an empirical study in developed and emerging economies. *The International Journal of Human Resource Management*, *25*(12), 1739–1762. https://doi.org/10.1080/09585192.2013.859164

Graefe, A., & Bohlken, N. (2020). Automated Journalism: A Meta-Analysis of Readers' Perceptions of Human-Written in Comparison to Automated News. *Media and Communication*, *8*(3), 50–59. https://doi.org/10.17645/mac.v8i3.3019

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2016). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, *19*(5), 595–610. https://doi.org/10.1177/1464884916641269

Grégoire, Y., & Fisher, R. J. (2008). Customer betrayal and retaliation: when your best customers become your worst enemies. *Journal of the Academy of Marketing Science*, *36*(2), 247–261. https://doi.org/10.1007/s11747-007-0054-0

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Hansen, M., Roca-Sales, M., Keegan, J., & King, G. (2017). *Artificial Intelligence: Practice and Implications for Journalism.* Columbia Journalism School, Columbia.

Hayes, A. F. (2018). *Introduction to Mediation, Moderation and Conditional Process Analysis. A Regression-Based Approach* (2nd edition). New York: The Guilford Press.

Highhouse, S., Lievens, F., & Sinar, E. F. (2003). Measuring Attraction to Organizations. *Educational and Psychological Measurement*, *63*(6), 986–1001. https://doi.org/10.1177/0013164403258403

Huang, M.-H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kietzmann, J., Paschen, J., & Treen, E. (2018). Artificial Intelligence in Advertising. *Journal of Advertising Research*, *58*(3), 263–267. https://doi.org/10.2501/JAR-2018-035

Kissel, P., & Büttgen, M. (2015). Using social media to communicate employer brand identity: The impact on corporate image and employer attractiveness. *Journal of Brand Management*, *22*(9), 755–777. https://doi.org/10.1057/bm.2015.42

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, *114*, 106553. https://doi.org/10.1016/j.chb.2020.106553

Lobosco, M. (2019). Here Are the Top Companies of 2019. Retrieved from https://business.linkedin.com/talent-solutions/blog/employer-brand/2019/linkedins-top-companies-2019

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science.* Advance online publication. https://doi.org/10.1287/mksc.2019.1192

Marr, B. (2019). Artificial Intelligence Can Now Write Amazing Content—What Does That Mean For Humans? Retrieved from https://www.forbes.com/sites/bernardmarr/2019/03/29/artificial-intelligence-can-now-write-amazing-content-what-does-that-mean-for-humans

Montal, T., & Reich, Z. (2017). I, Robot. You, Journalist. Who is the Author? *Digital Journalism*, *5*(7), 829–849. https://doi.org/10.1080/21670811.2016.1209083

Mulvey, T., Esen, E., & Coombs, J. (2016). Using Social Media for Talent Acquisition, Recruitment and Screening. Retrieved from www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Documents/SHRM-Social-Media-Recruiting-Screening-2015.pdf

Oliver, R. L. (1980). A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research*, *17*(4), 460. https://doi.org/10.2307/3150499

Poels, K., & Dewitte, S. (2019). The Role of Emotions in Advertising: A Call to Action. *Journal of Advertising*, *48*(1), 81–90. https://doi.org/10.1080/00913367.2019.1579688

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Rampl, L. V., Opitz, C., Welpe, I. M., & Kenning, P. (2016). The role of emotions in decision-making on employer brands: insights from functional magnetic resonance imaging (fMRI). *Marketing Letters*, *27*(2), 361–374. https://doi.org/10.1007/s11002-014-9335-9

Rim, H., Park, Y. E., & Song, D. (2020). Watch out when expectancy is violated: An experiment of inconsistent CSR message cueing. *Journal of Marketing Communications*, *26*(4), 343–361. https://doi.org/10.1080/13527266.2018.1523216

Riquelme, I. P., Román, S., Cuestas, P. J., & Iacobucci, D. (2019). The Dark Side of Good Reputation and Loyalty in Online Retailing: When Trust Leads to Retaliation through Price Unfairness. *Journal of Interactive Marketing*, *47*, 35–52. https://doi.org/10.1016/j.intmar.2018.12.002

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics. *Journal of Database Management*, *31*(2), 74–87. https://doi.org/10.4018/JDM.2020040105

Solaiman, I., Clark, J., & Brundlage, M. (2019). GPT-2: 1.5B Release. Retrieved from https://openai.com/blog/gpt-2-1-5b-release/

Thurman, N., Dörr, K., & Kunert, J. (2017). When Reporters Get Hands-on with Robo-Writing. *Digital Journalism*, *5*(10), 1240–1259. https://doi.org/10.1080/21670811.2017.1289819

Van Dalen, A. (2012). THE ALGORITHMS BEHIND THE HEADLINES. *Journalism Practice*, *6*(5-6), 648–658. https://doi.org/10.1080/17512786.2012.667268

Van der Kaa, H. A., & Kramer, E. J. (2014). Journalist versus news consumer: The perceived credibility of machine written news. *Proceedings of the Computation+Journalism Conference*.

Van Esch, P., Stewart Black, J., Franklin, D., & Harder, M. (2020). AI-enabled biometrics in recruiting: Insights from marketers for managers. *Australasian Marketing Journal (AMJ).* Advance online publication. https://doi.org/10.1016/j.ausmj.2020.04.003

Waddell, T. F. (2018). A Robot Wrote This? *Digital Journalism*, *6*(2), 236–255. https://doi.org/10.1080/21670811.2017.1384319

Wagner, T., Hennig-Thurau, T., & Rudolph, T. (2009). Does Customer Demotion Jeopardize Loyalty? *Journal of Marketing*, *73*(3), 69–85. https://doi.org/10.1509/jmkg.73.3.69

Wölker, A., & Powell, T. E. (2018). Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*, 146488491875707. https://doi.org/10.1177/1464884918757072

Wu, Y. (2019). Is Automated Journalistic Writing Less Biased? An Experimental Test of Auto-Written and Human-Written News Stories. *Journalism Practice*, *14*(8), 1008–1028. https://doi.org/10.1080/17512786.2019.1682940

Figure 1. Conceptual model

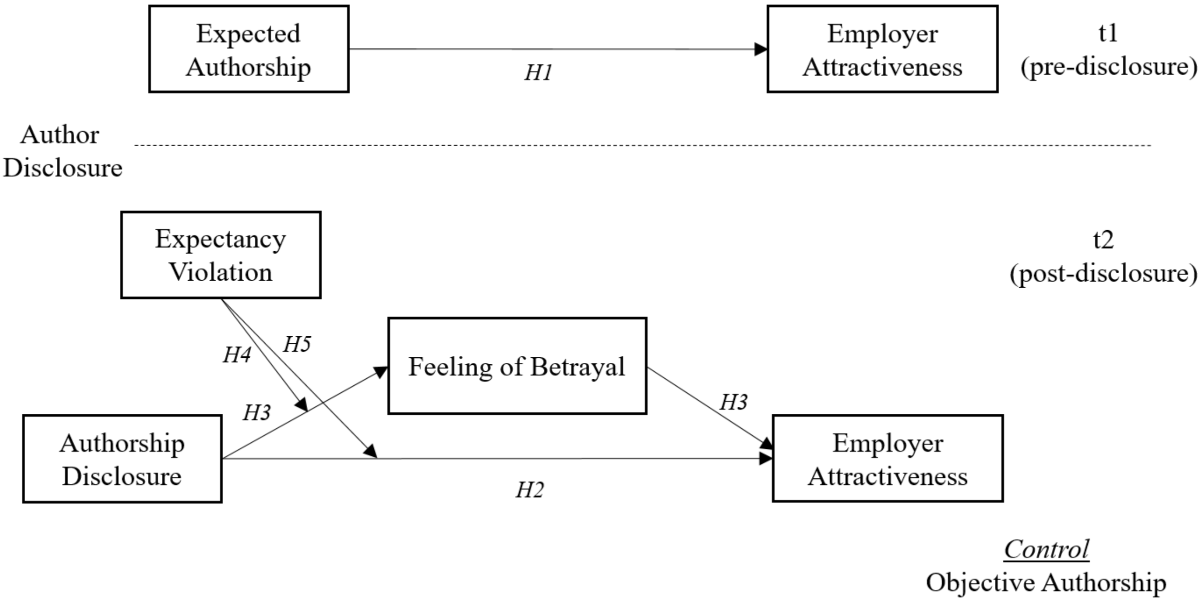


Figure 2. Effect of Author Disclosure on Employer Attractiveness ratings

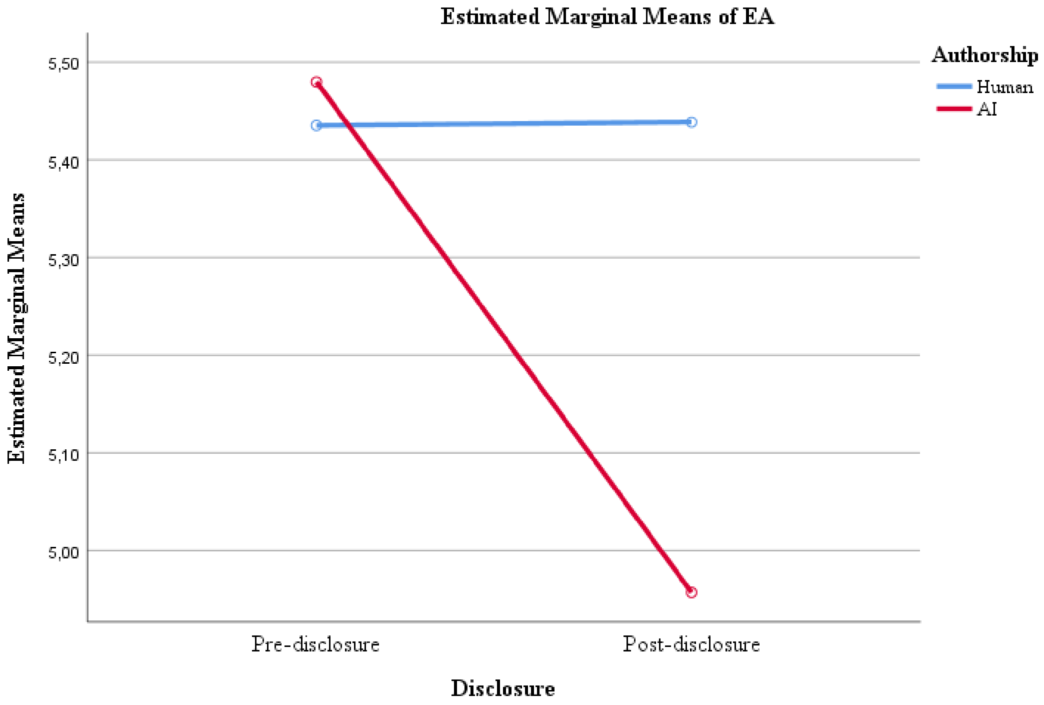


Figure 3. Effect of Author Expectation and Author Disclosure on Employer Attractiveness

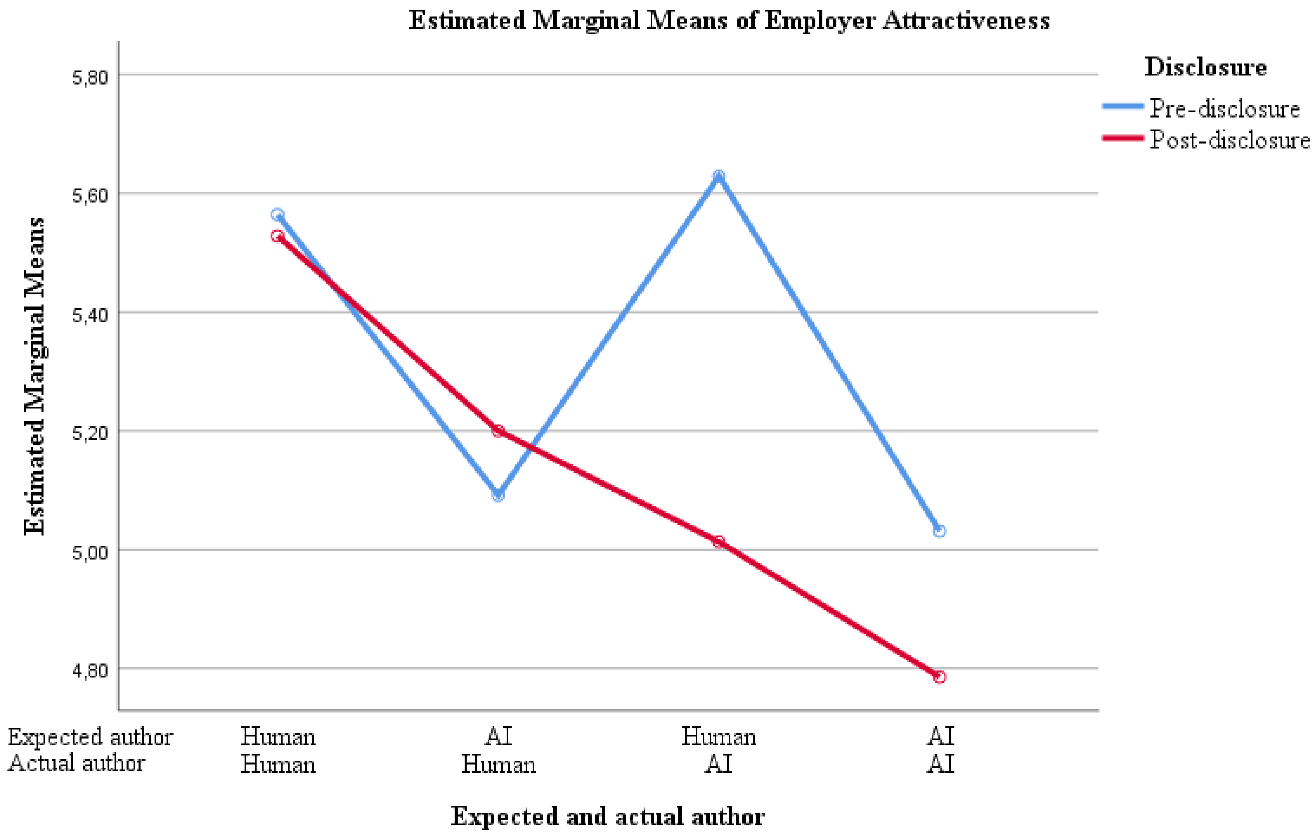


Table 1. Overview of Prior Research regarding AI created Content

Study	Type of Texts	Dependent Variable(s)	Mediator(s)/ Moderator(s)	Authorship disclosed?	Main findings
Wu (2019)	Sports, Politics, Finance reports	Perceived objectivity, Message and Medium Credibility, Bias, Quality	-	Y and N	AI (vs. human) written texts are evaluated as more objective, more credible and less biased. AI-texts are more credible when source and affiliation are not mentioned. Differences between news types exist.
Wölker and Powell (2018)	Sports and Finance reports	Source Credibility, News selection	Message Credibility (n.s.)	Y	AI (vs. human) written and combined content are evaluated equally credible. No mediating effect of message credibility on news selection.
Waddell (2018)	Weather, Stock market, Science news	News Credibility, Newsworthiness, Quality, Representativeness	Source anthropomorphism, machine heuristics, Message expectancies	Y	News attributed to AI (vs. human) is perceived as less credible, effect is mediated by source anthropomorphism and subsequently negative expectancies (e.g., disappointment when reading).
Graefe et al. (2016)	Sports and Finance reports	Credibility, Readability, Journalistic Expertise	-	Y	Articles declared as human (vs. AI) were favored. Actual AI-texts were rated as more credible and with higher expertise, but less readable.
Van der Kaa and Krahmer (2014)	Sports and finance reports	Expertise, Trustworthiness	-	Y	Differences in perceived expertise and trustworthiness were small but AI was rated more positive higher by news readers, the human author was rated more positive by journalist participants.
Clerwall (2014)	Sports report	Credibility, Readability	-	N	AI (vs. human) written news were rated as more credible but less readable, participants were unable to identify article source

Table 1. Overview of Prior Research regarding AI created Content (continued)

Study	**Type of Texts**	**Dependent Variable(s)**	**Mediator(s)/ Moderator(s)**	**Authorship disclosed?**	**Main findings**
Our Study	Firm image text on social media	Employer Attractiveness	Feeling of betrayal Author expectancy violation	Y and N	When source is unknown, AI and human-written texts are rated equally. AI (vs. human) disclosure reduces employer attractiveness, fully mediated by feelings of betrayal. Author expectancy violations (vs. match) moderate the effect on employer attractiveness.

Note. Y = Yes, N = No, n.s. = not significant.

Table 2. Robustness check results comparing AI text with Human text and AI text with human as declared author

	Betrayal	Employer Attractiveness (Pre-disclosure)	Employer Attractiveness (Post-disclosure)
AI text	4.14 (.14)	5.48 (.07)	4.96 (.81)
Human text	2.92 (.14)	5.46 (.07)	5.47 (.82)
Fake human (AI text declared as human text)	2.98 (.12)	5.48 (.07)	5.52 (.73)
Pairwise comparisons			
AI vs. Human text	***	n.s.	***
AI vs. Fake Human	***	n.s.	***
Human vs. Fake Human	n.s.	n.s.	n.s.

Note: *** p < .001, n.s. = not significant. Mean ratings based on 7-point scales, standard error in parentheses.

Appendix 1. Texts from LinkedIn – Original (human written) and AI-generated version

Original version

If you're excited about the innovations of tomorrow, we want to hear from you. There are opportunities to join us in almost every country in the world and you'll join thousands of the smartest minds, working together to build a better future. We recruit people from all walks of life: it doesn't matter the gender, age, ethnicity, sexual orientation or disability. Our aim is to find people from around the globe who can contribute to world-changing technology. Can you help us create what's next?

By reimagining the world around us, creating solutions to the most challenging problems and making infrastructure more intuitive, we're making the world a smarter place. Through electrification, digitalization and automation, we're changing the cities you live in, the places you work, the energy that powers your home, and the way the products you love are manufactured. It's innovations like these that inspire over 351,000 people at our company to come and work together every day.

With smart thinking and creativity, we're making a difference to the world we live in every day. We've reduced congestion in cities by 20% using intelligent real-time traffic data; we've built offices using a 3D printer; we've created software that communicates with turbines in the middle of the ocean; and we helped NASA develop the Curiosity rover and land it safely on Mars. By working with like-minded companies around the globe, we can create innovations that contribute to society's success.

AI-generated version

At our company, our purpose is to transform global technology into smart solutions that affect and benefit people. Whether that means building smarter infrastructure to support economic growth, providing clean energy, or enhancing security through encryption, smart solutions will allow us to solve tomorrow's challenges while creating new opportunities. Our employees share our mission to improve the world. They are passionate about technology and the things that make the world a better place.

Our philosophy is built around the following: Innovation comes from collaboration between individuals, teams, and companies. Innovation thrives when people are empowered to solve real world problems with their unique expertise and insight. People who can share their passion with others bring new ideas and solutions to life. Our culture encourages employees to use their skills to transform the world around them by designing innovative products and services for the future.

At our company, our people have a common goal: we will use our unique talents to create a better future for everyone. We believe that working together means we can do things that no one else can. Together, we are stronger. Together we innovate. Together we win. Learn more about our culture and how we help empower the world. We are an equal opportunity employer, and are committed to providing an environment that fosters diversity and inclusion for all individuals, regardless of race, gender, age, nationality, sexual orientation, religion, disability, or any other status, condition, or belief.

Appendix 2. Scale items and statistics

Construct name and items		Standardized loading: Pre-/Post-disclosure	
Feeling of Betrayal	(α = .93; CR = .96; AVE = .93)		
• I feel fooled by the company.			.96
• I feel tricked by the company.			.96
Employer attractiveness	(Pre-disclosure: α = .90; CR = .88; AVE = .64) (Post-disclosure: α = .92; CR = .90; AVE = .68)		
• This company is attractive to me as a place for employment		.82	.85
• For me, this company would be a good place to work		.80	.85
• I would not be interested in this company except as a last resort		(dropped)	
• I am interested in learning more about this company		.76	.75
• A job at this company is very appealing to me		.82	.84

Appendix 3. Constructs

Latent Construct Means, Standard Deviations, and Correlations

Construct	*M*	*SD*	1	2	3
1. Betrayal	3.51	2.18	**.96**		
2. Employer Attractiveness (Pre-disclosure)	5.46	1.18	.17	**.80**	
3. Employer Attractiveness (Post disclosure)	5.20	1.40	-.02	.73	**.83**

Note. Values on the diagonal are square roots of AVE, values below the diagonal are correlations. [a] not calculated for single-item measure

Paper 4

Using Human-AI Collaborative Authorships to Escape Algorithm Aversion in Content Marketing

Status: under review after revise-and-resubmit at the journal *Review of Managerial Science*

Co-Authors: Jan Freidank and Alexander Haas

Own Share: 80%

Using Human-AI Collaborative Authorships to Escape Algorithm Aversion in Content Marketing

Abstract

Although Artificial intelligence (AI) can offer significant business benefits, many consumers have negative perceptions of AI, leading to negative reactions when companies act ethically and disclose its use. Based on the pervasive example of content creation (e.g., via tools like ChatGPT), this research examines the potential for human-AI collaboration to preserve consumers' message credibility judgments and attitudes towards the company. The study compares two distinct forms of human-AI collaboration, namely human authorship with AI support or AI authorship with human control, with traditional human authorship or full automation. Building on the compensatory control theory and the algorithm aversion concept, the study evaluates whether disclosing a high human input share (without explicit control) or human control over AI (with lower human input share) can mitigate negative consumer reactions. Moreover, this paper evaluates the moderating role of consumers' perceived morality of companies' AI use. Results from two experiments in different contexts reveal that human-AI collaboration can alleviate negative consumer responses, but only when the collaboration indicates human control over AI. Furthermore, the effects of content authorship depend on consumers' moral acceptance of a company's AI use. AI authorship forms without human control lead to more negative consumer responses in case of low perceived morality (and no effects in case of high morality), whereas messages from AI with human control were not perceived differently to human authorship, irrespective of the morality level. These findings provide guidance for managers on how to effectively integrate human-AI collaboration into consumer-facing applications and advises to take consumers' ethical concerns into account.

Keywords: Artificial intelligence, Human-AI collaboration, AI augmentation, AI ethics, content creation, algorithm aversion

1. Introduction

Artificial intelligence (AI) currently reshapes business and marketing strategies, as companies increasingly rely on the use of AI systems to create product recommendations, offer driverless vehicles, analyze customer demand or give medical advice (Longoni et al., 2019; Osburg et al., 2022; Yeomans et al., 2019). As the development of AI continues to advance rapidly, the importance of studying human perception and responses to AI tools increases (Kaplan & Haenlein, 2019). Despite an increased diffusion of AI-tools, several researchers found that consumers tend to have an *algorithm aversion*, which refers to a negative perception of AI and an unwillingness to trust in and use algorithms although they often outperform humans (Burton et al., 2020; Castelo et al., 2019; Dietvorst et al., 2015, 2016; Yeomans et al., 2019). For instance, individuals are frightened that AI will attain too much power and get beyond human control, or companies might misuse this power to act unethically (Alfonseca et al., 2021; Burton et al., 2020; Siau & Wang, 2020). These negative perceptions are critical for companies which use AI, as transparent AI declaration as ethical principle will become a legal obligation in the near future for many societies. For instance, the European AI Act mandates companies to disclose AI-generated data and is due to come into force by the end of 2023 (European Parliament, 2023), and California has already passed a bot bill to legally enforce transparent AI declaration (Bots: Disclosure, 2018/SB-1001).

Thus, managers using AI-tools in consumer-facing applications are increasingly confronted with the question of how to integrate AI transparently without suffering from negative consumer responses. As a viable, but under-researched option, companies could use human-AI collaboration to leverage the efficiency of AI while maintaining the positive effects of a human (Huang & Rust, 2022; Zanzotto, 2019). According to scholars, many current AI-driven applications still rely on (some) human input or control at some point anyhow (Huang & Rust, 2022). In addition, it has been suggested that collaborative work between humans and AI increases trust in AI systems and managers' perceptions of empowerment (i.e. the ability to adapt or change) (Schleith et al., 2022). In the related field of AI-controlled machines (e.g., robots in service or elderly care), such collaboration between humans and AI are already a common phenomenon (Larivière et al., 2017). More generally, this topic of human-AI cooperation as "Intelligence Augmentation" has received increasing scholarly attention lately (Hassani et al., 2020; M. Langer & Landers, 2021; Raftopoulos et al., 2023; Zhou et al., 2021).

Among the manifold options to use AI for business and marketing, this study focuses on the use case of AI-powered content generation tools. Applications such as OpenAI's ChatGPT have seen tremendous interest from both business and academia. For example, ChatGPT has attracted more than 1 million users in five days and is expected to transform many tasks in marketing, information systems, or journalism, among others. It is even seen as a serious threat to Google by providing more human-like responses and a smoother experience (Olson, 2022). AI is increasingly able to create compelling content that cannot be distinguished from human-authored texts any more (Köbis & Mossink, 2021; Waddell, 2018), and scholars identified content generation as a key application area of AI in marketing and other business fields (Dwivedi et al., 2023; Graefe & Bohlken, 2020; Kahnt, 2019). A recent report from McKinsey acknowledged this and described that content-generating AI-tools like ChatGPT could serve manifold business functions, including marketing and sales, finance and accounting, human resources, IT, risk and legal, among others (Chui et al., 2022).

Despite a growing body of research regarding human-AI collaboration in various fields (see Table 1 for a literature overview), empirical studies regarding the use and declaration of human-AI collaboration for content creation remain limited (Huang & Rust, 2022). Most studies related to AI and content creation come from the field of journalism and compare message evaluations of AI authors versus a human (Graefe et al., 2016; Waddell, 2018; Wu, 2019). Yet, a complementary human-AI cooperation is expected to be more realistic in content creation than a human replacement in the near future (Bailer et al., 2022). This prognosis is shared by Huang and Rust (2022), who assert that AI will first augment and (only) later replace humans at a given level of intelligence. In addition, scholars propose that the combination of human and AI will lead to higher value and competitive advantage compared to human replacement, e.g., in education, medicine, business, science and others. In particular, a collaboration is suggested to merge the respective strengths of humans (e.g., build rapport and trust with customers or patients; create a personal touch, address ambiguities) and of AI (e.g., analyze large datasets, find anomalies, optimize materials and layout, do mundane or repetitive work) (Paschen et al., 2020; Zhou et al., 2021).

In the business and marketing context, there are to date no studies explicitly dedicated to the impacts of collaborative authorships. Two initial studies are present in the adjacent field of journalism. Waddell (2019) found that collaborative human-AI authorship (vs. human author) have both positive and negative effects on message credibility. In particular, the collaboration increased credibility perceptions via reduced feelings of bias, but was also perceived as less credible than a human due to lower source anthropomorphism. Wölker and Powell (2018) assessed that the credibility perceptions of human and collaboratively created content were not significantly different. Although the differences were quite small, the mean credibility scores of human-AI authorship ranged between a sole AI and a sole human author.

In sum, effects of human-AI collaboration remains under-researched, particularly for content creation as highly prominent example for AI-integration in business (Chui et al., 2022). This is surprising, given the high potential to create efficiencies and the high performance level of modern text-generating tools such as ChatGPT (Dwivedi et al., 2023) in combination with the challenges due to legislative requirements for AI transparency (European Parliament, 2023) and individuals' algorithm aversion. To fill this gap, this research first aims to identify if human-AI collaboration could work as an escape from negative consequences of AI disclosure.

Moreover, human-AI collaboration can be designed in different ways and constitutes a continuum between the end-points of a sole human actor and sole AI. Following Huang and Rust's (2022) framework of collaborative AI, human-AI cooperation regularly follows a stepwise pattern: Due to the permanent development of AI, AI starts as augmentation and support for humans, and later could replace humans and fulfill the task autonomously. However, as a third format in between, several scholars acknowledge that AI might perform the task under the surveillance and control of a human (Longoni et al., 2019; Nyholm, 2022; Osburg et al., 2022). Related to content creation, Bailer et al. (2022) support these stepwise collaboration forms and distinguish between AI supporting a human (i.e., labeled as “AI in the loop of human intelligence“) vs. AI task take-over with human control (i.e., labeled as “human in the loop of AI”). Although scholars have acknowledged these different collaboration formats, research currently lacks empirical evidence regarding the impact of these different forms of collaboration between humans and AI. Therefore, the second aim of this study is to evaluate the effectiveness of the two distinct types of human-AI collaboration in terms of consumer responses to the message and the company.

Based on two experimental studies in different contexts, this study compares the two distinct human-AI collaborative forms (1) human authorship with AI support and (2) AI authorship with human control (Bailer et al., 2022) - with full AI automation and the traditional form of human authorship. Thereby, the first collaboration form (i.e., “human authorship with AI support”) is suggested to declare a relatively high human input share but does not express clear human control, whereas the second form (i.e., “AI authorship with human control”) declares a lower human input share with an explicit human control function With this research, our study answers the calls of several scholars to calls to assess consumer perceptions and responses of human-AI collaborations (Wölker & Powell, 2018) and insights about the optimal design of such collaborative forms (Huang & Rust, 2022; Zhou et al., 2021).

To hypothesize and assess the effects of the two formats (vs. a traditional human author) on consumer responses and company-related outcomes, we build on the aforementioned concept of *algorithm aversion* (Burton et al., 2020) and the *compensatory control theory* (Landau et al., 2015) . This theory describes that individuals “respond to events and cognitions that reduce control with compensatory strategies" to regain their control (Landau et al., 2015, p. 694). Regarding (negative) consequences of authorship declaration, we focus on message credibility and its role as a mediating variable between authorship declaration and company evaluation. Ultimately, human-AI collaboration aims to reach human levels of message credibility, as this level was traditionally given before AI integration. This research thus contributes to gain a better understanding of human-AI collaboration effects for consumer-facing applications.

The increasing use of AI in business and marketing has also sparked discussions about companies' ethical behavior and consumers' perceptions of morality (Cremer & Kasparov, 2021; Hagendorff, 2020; Siau & Wang, 2020; Wirtz et al., 2022). For instance, Wirtz et al. (2022) have established the term of "corporate digital responsibility" and refer it to a company's "ethical, fair, and protective use of data and technology when engaging with customers within their digital service ecosystem" (p. 1). Further, scholars clearly acknowledged that the use of AI in marketing can lead to substantial ethical challenges, including consumer discrimination, privacy issues, intransparent information or loss of control (Cremer & Kasparov, 2021; Hagendorff, 2020; Siau & Wang, 2020; Wirtz et al., 2022).

In addition, integrating AI into marketing also creates the need for different stakeholders to balance trade-offs. For instance, consumers have to decide between highly personalized offers and privacy concerns (Davenport et al., 2020; Paschen et al., 2020), and companies have to decide between customer value maximization through customer priorization and service discrimination (Libai et al., 2020). AI use for highly personalized content and advertising is also criticized to cause addictive behaviors and public health issues (Daza & Ilozumba, 2022).

When people are confronted with AI, different moral judgments could be elicited which are expected to affect their message credibility evaluations. Therefore, our study introduces 'perceived morality of companies' AI use' as an influencing factor that moderates the effects of the different authorship options. Thus, we extend findings to the new field of AI ethics and link it to marketing strategy (Siau & Wang, 2020).

For managers, this study offers a solution to escape the dilemma between ethical (and upcoming legal) misconduct by hiding AI use and negative consumer reactions to transparent AI use. The findings show that a collaboration between humans and AI can protect a company's reputation, as long as the collaboration clearly indicates human control over AI (vs. an AI-supported collaboration). Furthermore, this effect remained robust even when morality perceptions of using AI were low, providing managers with the opportunity to implement an uncritical, efficient, and highly effective collaboration between humans and AI.

In the following sections, we review the literature on AI performance and consumers' perception of AI and discuss causes of negative responses to AI use. We then present human-AI collaboration as a possible solution to mitigate these negative reactions and propose different evaluation schemes that consumers might use to judge such human-AI collaboration. Next, two experimental studies test the hypothesized effects, followed by a discussion of the major findings, implications for theory and practice as well as limitations that provide future research opportunities.

2. Theoretical background and hypotheses development

2.1 Performance and perceptions of AI

Related to consumer research, AI can be defined as "any machine that uses any kind of algorithm or statistical model to perform perceptual, cognitive, and conversational functions typical of the human mind" (Longoni et al., 2019, p. 630). Parallel to other disciplines, the precision and effectiveness of AI in content creation is rapidly developing (Dwivedi et al., 2023). Early studies on the perception of AI-generated content mainly focused on content quality, such as the correct wording, grammar, or coherent structure of the information presented (Mellish et al., 2006). Quality and precision of AI are rising drastically and AI content is often not distinguishable from human written content (Köbis & Mossink, 2021). A recent meta-study from Graefe and Bohlken (2020) showed, that AI-based texts achieved comparable evaluations to human-written texts in various studies – as long as content authorship was hidden. However, when the use of AI is transparent, consumers were found to react differently.

Transparent AI triggers Algorithm aversion

When companies transparently declare their use of AI, scholars widely observed an *algorithm aversion,* implying that consumers are reluctant to use AI (compared to humans) (Castelo et al., 2019; Dietvorst et al., 2015). This phenomenon was found in various instances, including product- and service-recommendations (Longoni & Cian, 2022; Wien & Peluso, 2021), performance-related forecasts (Dietvorst et al., 2015), and financial or medical advice (Longoni et al., 2019; Önkal et al., 2009). A systematic literature review of Burton et al. (2020) revealed that algorithm aversion could be consistently documented since the 1950's and can be attributed to several causes, such as consumers' (mis-)understanding and expectations of AI or their reluctance to delegate decision-making autonomy to AI. As one of the observed studies, Dietvorst et al. (2015) showed that people consistently choose a human forecaster (vs. an algorithm) after seeing them performing a predictive task. People were also found to trust medical advice less when it is provided by AI (vs. a human doctor), even though AI could often outperform human doctors (Longoni et al., 2019). Similarly, an AI-driven digital agent (i.e., chatbot) was equally effective as a competent human sales agent in terms of conversion rates – but only as long as the chatbot's identity was hidden. By disclosing the AI identity, the purchase rate dropped by over 75% (Luo et al., 2019). A recent study of Longoni and Cian (2022) showed that product and service attributes (i.e., hedonic or utilitarian contexts) determine whether people prefer AI or human advice, and thus act as a boundary condition for the algorithm aversion effect.

Corresponding to these findings, algorithm aversion could be also found related to AI content creation., Individuals often assigned higher ratings regarding credibility, readability or quality to human (vs. AI) generated content when authorship was transparent (Graefe & Bohlken, 2020; Waddell, 2018). Graefe and Bohlken (2020) showed that these ratings were even made regardless of the actual source. That means, despite an identical text, the assignment of an AI (vs. human) authorship systematically leads to more negative ratings.

In sum, these negative reactions seem to be driven mainly by a low subjective AI source credibility rather than a lack of objective AI quality. Essentially, source credibility could be defined as “qualities of an information source which cause what it says to be believable” (West, 1994, p. 159). According to the source credibility theory (Hovland et al., 1953), individuals are more likely to be persuaded when the source is evaluated as credible (i.e., expertful and trustworthy). Manifold studies throughout the last decades support this proposition (for an overview, see Ismagilova et al. (2020)). More (vs. less) credible sources were found to create favorable outcomes including enhanced message evaluations, attitudes and behavioral intentions. For instance, high source credibility leads to higher brand trust or purchase intentions (Harmon & Coney, 1982; Luo et al., 2019; Ohanian, 1990; Visentin et al., 2019). Moreover, source credibility significantly increases message credibility perceptions (Ismagilova et al., 2020; Visentin et al., 2019) and thus, even the same content could be perceived differently due to different sources. Essentially, message credibility refers to “an individual’s judgment of the veracity of the content of communication” (Appelman & Sundar, 2016, p. 63).

While most studies supported this algorithm aversion (see Graefe & Bohlken, 2020), some studies found no effect or even a positive effect of AI authorship on perceived content credibility, e.g., in sports news (Wölker & Powell, 2018). Thus, although algorithm aversion dominates human perceptions of AI, the effect was not fully consistent throughout content topics or AI tasks.

As possible reasons for this aversion, scholars asserted that humans rated AI generally as less trustworthy, less empathetic and less competent (Chan-Olmsted, 2019; Luo et al., 2019). For instance, in a sales call, humans replied more concise and purchased less from an AI-driven chatbot because they perceived the chatbot as less knowledgeable and less empathetic than a human salesperson (Luo et al., 2019). Similarly, Castelo et al. (2019) showed that consumers assume that AI is incapable to successfully complete subjective tasks, leading to lower trust and reliance on AI. Second, humans seemed to expect more perfect results from an AI than from a human, and seeing AI err led to lower confidence towards the AI and an AI rejection for further tasks (Dietvorst et al., 2015). Third, many processes of AI, such as machine learning, are hard to explain – even for their creators, and thus are often considered as inherently intransparent or as "black box" (Siau & Wang, 2020). This deficit of understanding AI creates information asymmetries and fuels fears and distrust (Puntoni et al., 2021).

Transparent AI triggers perceived loss of control

In addition to that, many people fear that AI could take over control in several domains or threaten human jobs (Huang & Rust, 2022). These feelings are not unjustified. When AI takes over a task, it often replaces human intelligence and inevitably takes away human control and jobs as long-term consequence (e.g., autonomous cars replace taxi drivers (Frey & Osborne, 2017; Huang & Rust, 2022; Osburg et al., 2022) or AI agents replace journalists (Yerushalmy, 2023)). Scholars consent that already "the mere recognition of AI's capability to act as a substitute for human labor can be psychologically threatening" (Puntoni et al., 2021, p. 140).

The desire for control is an essential human need and refers to people's desire to be able to manage processes and outcomes of events in life (Burton et al., 2020; Chen et al., 2017; Puntoni et al., 2021). Herein, control refers to the ability to influence outcomes in one's environment (Skinner, 1996). When this need for control is threatened or remains unmet, people experience negative affect, including discomfort, frustration, demotivation, and helplessness, and respond with negative behavior such as moral outrage or reactance (Chen et al., 2017; Landau et al., 2015; Puntoni et al., 2021). Furthermore, according to the compensatory control theory (Landau et al., 2015), individuals who experience a reduced level of control respond with compensatory strategies to restore their perceived control. As traditional strategy, people bolster their personal agency, which is their belief that they possess the resources needed to perform a specific action (E. J. Langer, 1975). According to a recent literature review of Cutright and Wu (2023), perceptions of low control shape consumers behavior either by motivating them to look for a sense of control and order in their consumption environment; or by motivating them to use consumption as a function to regain control. A growing body of literature examines how product acquisition satisfies consumers' need for control (Billore & Anisimova, 2021; Chen et al., 2017; Cutright & Wu, 2023). For example, in the recent COVID 19 pandemic, people perceived a lack of control over their lives (Cutright & Wu, 2023) and were found to engage in 'panic buying', where they irrationally purchased high quantities of household or medical goods to regain a sense of control over the situation (Billore & Anisimova, 2021). Related to algorithms and forecasts, people were found be less reluctant to use AI when they could modify the results because they feel more in control, even when this control means only a slight modification (Dietvorst et al., 2016).

When it comes to content marketing, people might be afraid that AI becomes so sophisticated that they will not be able to distinguish an AI from a human author (which would be supported by research results like the study from Köbis & Mossink, 2021), resulting in a lack of control over the message provider. This, in turn, creates the fear that companies might use AI and manipulate consumers' activities and perceptions (Jobin et al., 2019). For instance, consumers feel uncertain whether a content is genuine human or not – and who controls it (Graefe & Bohlken, 2020).

People do not only rely on themselves but also on other humans to restore control. Therefore, a further strategy mentioned in the compensatory control theory is the so-called secondary control, which is a person's belief to have access to an external agent who possesses a desired or needed ability (Landau et al., 2015). That means, a person or institution outside of one's self can influence personally important outcomes and increase the chances to achieve one's goals (Friesen et al., 2014; Kay et al., 2008; Landau et al., 2015). Scholars showed that when people feel a lack of control, they rely stronger on other entities which provide clear rules and structures and thus, satisfy their desire for order and control (Friesen et al., 2014; Kay et al., 2008). For instance, individuals were more supportive of hierarchies in the workplace (and favored hierarchy-enhancing jobs) when their sense of control was threatened (Friesen et al., 2014). Similarly, a study with people from 67 nations showed that lower perceived control is strongly correlated with higher support of governmental control (Kay et al., 2008). We adopted this concept of secondary control to our research design as control is exerted by the external agent – the human author.

2.2 Human-AI collaboration as possible escape to negative reactions to AI

One possible, but under-researched, solution to mitigate the negative consequences of AI authorship lies in the collaboration of humans and AI, meaning that "AI systems work jointly with humans like teammates or partners to solve problems" (Lai et al., 2021, p. 390). For various marketing tasks, scholars consent that AI and humans could collaborate in manifold ways to use the respective strengths of humans and AI (Huang & Rust, 2022; Raftopoulos et al., 2023; Zhou et al., 2021). For instance, human-AI collaboration can support healthcare professionals (Lai et al., 2021), general management (Sowa et al., 2021) or data scientists (Wang et al., 2022).

Humans could also collaborate with AI in advertising (Vakratsas & Wang, 2021), marketing planning and strategy (Ameen et al., 2022), or jointly deliver customer service (Wirtz et al., 2018). AI could also augment salespersons' capabilities in every stage of the sales process (Davenport et al., 2020; Paschen et al., 2020). For example, AI could detect unmentioned complaints with the help of automated customer's voice analysis and a human salesperson could follow up on this (Davenport et al., 2020). AI could also predict leads and personalize content, whereas the human could verify leads and link them to the business context (Paschen et al., 2020). Next to this, AI could support user experience evaluations (e.g., by identifying issues in usability test videos) to enhance user engagement and sales (Fan et al., 2022).

In general, scholars suggest that AI is more effective when it augments (vs. replaces) human marketing managers (Davenport et al., 2020), and empirical evidence showed that engaging in collaborative tasks with an AI-driven robot increased consumers' rapport, cooperation and engagement levels (Seo et al., 2018). In the field of healthcare provision, researchers revealed that human-AI collaboration represents a solution to escape consumers' resistance to AI (medical) services (Longoni et al., 2019; Tobia et al., 2021).

An analysis of human-AI collaboration effects in content marketing is missing. However, scholars in the related field of journalism and news production have started to evaluate this and labeled it "hybrid" or "tandem" authorship. Several authors draw optimistic scenarios where AI could be integrated into journalistic work, and AI and journalists could reach a state of cooperation instead of cannibalization (Graefe & Bohlken, 2020; Graefe et al., 2016; Wölker & Powell, 2018). Supporting that, Waddell (2019) asserts pragmatically that many current AI systems in journalism still need some human input anyhow, therefore, mentioning both human and AI as cooperative authors is recommended. Empirically, Wölker and Powell (2018) show that a human-AI collaboration for largely standardized sports and finance reports is perceived as equally credible source as a human author, and the collaboration did not lead to lower news selection (Wölker & Powell, 2018). These scholars assume that this might be rooted either in the perception of an AI as more objective author or in initially low expectations toward AI authorship.

Several studies from diverse fields showed that integrating humans into AI tasks is reducing their initial algorithm aversion (Burton et al., 2020; Dietvorst et al., 2016; Tobia et al., 2021). This phenomenon is sometimes also described as „human-in-the-loop“ decision making, and describes a system where the human user could “semi-supervise the algorithm by having opportunities to intervene, provide input, and have the final say” (Burton et al., 2020, p. 224). Following this explanation, Human-AI collaboration can be evaluated based on different schemes. As cues for the message credibility evaluation, people could either focus on whether human or AI provided the major part of input, or the level of perceived human authority and control over AI (Burton et al., 2020; Dietvorst et al., 2016).

The effects of input share and control

A traditional criterion to evaluate content of two authors is to base the decision on the particular workload or input each author provided. For instance, in academic content with cooperative authorship, the authorship order reflects the level of contribution and input share (Newman & Jones, 2006). Given the tendency of people’s AI avoidance (Longoni et al., 2019; Luo et al., 2019), higher input share of a human (AI) author is expected to be perceived as more positive (negative). Thus, a higher level of AI input share is expected to reduce message credibility evaluations because people generally rate AI as a less credible source (Luo et al., 2019).

Next to this, people could also evaluate a human-AI collaboration based on the perceived level of human authority and control over AI in the content creation process. As mentioned above (see 2.2), the use of AI as an autonomous system deprives people’s sense of control over processes and outcomes (Huang & Rust, 2022; Osburg et al., 2022). To counteract this, humans act as supervisors in many processes where AI is used. For instance, humans supervise AI’s (semi-)autonomous steering of a car, or a human doctor controls AI’s medical advice (Longoni et al., 2019; Osburg et al., 2022).

Related to AI authorship in content creation, it is practically impossible for the readers to influence who writes the text or to verify the content’s truthfulness (i.e., objectivity and honesty) (Waddell, 2019). Instead, the reader has to rely on secondary control whenever possible – for instance to trust a human co-author or editor and to hand over the control or verification of the content to them.

In general, the desire to have or restore control over one's environment was found to be an innate human need and a quite strong motivator. For instance, when people's feeling of control is impaired, they react with strongly negative affect including anger, moral outrage or reactance (Puntoni et al., 2021). Longoni et al. (2019) find that people's resistance to use medical AI could be alleviated when AI supported a human who makes the final decision (i.e., is in control) instead of a sole AI service provision. These results support the effectiveness of the form "AI author controlled by a human".

In contrast, a high human share of input (as indicated in the form "Human supported by AI") is expected to be a less clear and powerful cue for the evaluation of message credibility. Particularly when the human input is not clearly visible and distinguishable from AI input (e.g., as mainly given in human-AI collaborative cases), people perceive a higher level of machine agency compared to human agency, and thus a lack of authority (Sundar, 2020). Moreover, without human control, individuals might perceive an increased risk of incorrect information (or action) from AI's input as no hierarchies and control functions are sought to be in place (Osburg et al., 2022). Thus, human control over AI is thought to have a stronger positive influence on message credibility perceptions than human input share. In particular, when human control is not specified, a high level of human input is not expected to reduce the negative impact of AI authorship (vs. a sole human authored message). However, when human control is stated, the perception of secondary control can mitigate the negative impact of AI authorship even with less human input. We hypothesize:

H1a. *Collaborative human-AI authorship types declaring high human input share but not declaring human control over AI lead to lower message credibility evaluations (vs. human authorship).*

H1b. *Collaborative human-AI authorship types declaring low human input share but also declaring human control over AI do not lead to different message credibility evaluations (vs. human authorship).*

Following persuasion research, message credibility affects how people make subsequent judgments about the message sending institution, such as companies or news agencies (Hovland et al., 1953). In particular, credible messages were found to increase consumers' trust and attitudes towards the message sender, and favorable behavioral intentions, including information adoption or purchase intentions (Ismagilova et al., 2020; Wölker & Powell, 2018). Thereby, a positive attitude towards the company refers to a readers' positive impression of the company, its reputation or image (Darke et al., 2008). We posit:

H2. *Stronger perceptions of message credibility lead to more positive attitudes towards the company.*

2.3 The moderating role of morality of AI use

Due to the increasing popularity of AI technologies, AI has gained a substantial impact on humans and society (Hagendorff, 2020). As a response to continuous calls from research, politics and even business, the new field of *AI ethics* as a part of applied ethics gains relevance and momentum (Hagendorff, 2020; Siau & Wang, 2020). Traditionally, ethics is defined as the "code of moral principles that sets standards as to what is good or bad, or right or wrong in one's conduct and thereby guides the behavior of a person or group" (Schermerhorn, 2002, p. 146). According to several scholars in the field of business ethics, ethics and morality are often used as synonyms in research as well as in public debate (see Fisher (2004) for an overview). For instance, according to (Shaw & Barry, 2014, p. 3)"almost everyone uses "ethical" and "moral" interchangeably to describe people we consider good and actions we consider right, and ''unethical'' and ''immoral'' to designate bad people and wrong actions."

Based on this understanding of (traditional) ethics, scholars refer to AI ethics as specific sub-dimension of ethics concerning all ethical issues which relate to AI, including issues originating from the development or from the use of AI (Daza & Ilozumba, 2022; Siau & Wang, 2020). As overarching goals, AI ethics should promote benefits for humans, foster social good ("beneficence"), and prevent any harmful consequences (non-maleficence) (Hermann, 2022; Jobin et al., 2019).

As many consumers were found to have moral concerns and reservations toward AI, discussions about the morality of companies' AI use are ongoing in different domains and consider multiple facets (Siau & Wang, 2020). Popular moral concerns are the lack of AI control, non-transparent AI processes ("black box"), discrimination, or low reliability of AI-created information (Jobin et al., 2019; Puntoni et al., 2021; Rai, 2020). Furthermore, scholars acknowledged possible morality issues when AI is integrated in consumer-facing applications because it could reduce consumer autonomy (Libai et al., 2020) and might be a highly manipulative system that could cause or support addictive user behavior (Daza & Ilozumba, 2022; Hermann, 2022). For example, AI could foster exhaustive social media usage through hyper-personalization and optimization of preferred content and ads, which increases marketing effectiveness but is also detrimental to public health (e.g., causing depression or anxiety) (Daza & Ilozumba, 2022). Finally, a recent study warned that the increased use of ChatGPT or related AI-driven technologies is supposed to create immense ethical issues, including a rising level of disinformation due to automated fake news, massive low-quality content creation and a more indirect communication between stakeholders in the society (Illia et al., 2023).

Nevertheless, the strength of these moral concerns related to AI technologies varies from person to person. In particular, some people were found to have a high technological affinity and are less worried about morality issues or possible downsides of AI use (Parasuraman & Colby, 2015; Puntoni et al., 2021). These individuals might mainly focus on the innovativeness of AI and have little concerns about moral violations related to their privacy or freedom in decision-making. In contrast, other consumers perceive a high risk and rather distrust AI. This group is more likely to believe that AI is employed to deceive them or take over control (Burton et al., 2020; Parasuraman & Colby, 2015). In general, moral judgements were found to influence consumers' perceptions and behavior (Finkel & Krämer, 2022; Schermerhorn, 2002; Siau & Wang, 2020). Research showed that perception of (non-) ethical behavior of a company is an important factor during the purchase decision process. Individuals rewarded a company's ethical behavior by showing a higher willingness to purchase and by paying higher prices for products (Creyer & Ross, 1997). Moreover, a recent study in the related field of humanoid robots revealed that consumers' morality perceptions positively influenced robot credibility attributions (Finkel & Krämer, 2022). Similarly, related to video news, positive morality judgments were found to lead to higher message credibility (Nelson & Park, 2015). Building on these results, we expect that moral judgements will also influence message credibility perceptions and downstream attitudes and behaviors. In particular, when people perceive companies' AI use as immoral (i.e. low morality), the use and declaration of authorship forms with AI involvement (i.e., AI or human-AI collaborative authorships) is sought to harm message credibility perceptions. In contrast, when people perceive companies' AI use as morally acceptable (i.e., high morality), the actual use of AI as sole author or co-author should not be an ethical issue. As these consumers exhibit lower moral objections to this kind of AI use, AI should also be perceived as a credible (co-)author, similar to a traditional human author (Creyer & Ross, 1997). In turn, high morality perceptions are expected to delete the negative effects of authorships on message credibility where AI is involved. Thus, we hypothesize:

> ***H3.*** *Perceived morality of AI use moderates the relationship between author type and message credibility: In case of low perceived morality of AI use, message credibility is lower for authorships where AI is involved than for human authorship, and there is no difference in message credibility across authorship types when perceived morality of AI use is high.*

Figure 1 depicts the conceptual model.

[Insert Figure 1 about here]

3. Empirical Studies

3.1. Study 1

Participants and Procedure

To examine the proposed causal relationships, we created an experiment and embedded it into an online survey (Hulland et al., 2018). In exchange for a small compensation ($ 0.75), participants (with a 95% approval rate in former tasks) were recruited from the platform Prolific. Prolific is one of the largest online platforms with over 130,000 participants and widely used in management research to conduct surveys or experiments. These platforms generally reach a more diverse population than traditional sampling methods and allow a quite rapid and inexpensive data collection (Gosling & Mason, 2015). To control for possible effects from a respondent's country of origin, we recruited participants with English as native language from the U.S. and UK. These countries were chosen as many AI-related studies are based on one of these Western countries and the pool of respondents was large enough to ensure a variety of participants.

After excluding participants who failed the attention check (i.e., "If you read this, please press button 4"), the final sample consisted of 243 participants (54.3 % female, M_{age} =35 years, SD_{age} = 18.29). As scenario, respondents were exposed to a product information website (i.e., depicting information about a jeans) from a fictitious clothing company (see Figure 3). The website was equal across all conditions, except for the author label. Respondents were randomly assigned to one of four experimental conditions, and read one of the following author descriptions: The text was created by (1) a human author (label: "Written by Mary Smith"), (2) a human author and AI support (label: “Written by Mary Smith supported by Artificial Intelligence”), (3) an AI author with human control (label: “Generated by Artificial Intelligence controlled by Mary Smith”, (4) an AI author (label: “Generated by Artificial Intelligence”).

Measures

After seeing the respective scenario, participants were asked to rate their perceived message credibility (Appelman & Sundar, 2016; Obermiller et al., 2005) with four items on a 7-point Likert scale (from 1 = “strongly disagree” to 7 = “strongly agree”). Furthermore, three items were used to assess respondents’ attitude towards the company (Darke et al., 2008). Next, we integrated an attention check item and evaluated the case realism with two items from Wagner et al. (2009), namely, “I believe that the described situation could happen in real life” and “I could imagine reading a text like the one presented earlier in real life”, $\alpha = 0.86$; M: 5.26, SD: 1.46). Finally, we asked for participants’ age, gender and education. No significant differences were found between the author groups regarding these three control variables (each $p > 0.1$), suggesting a successful randomization. All psychometric measures were above the recommended levels (see Table 2), indicating construct reliability and validity (Hulland et al., 2018).

As manipulation check, respondents were asked to estimate the share of human vs. AI input. Figure 2 illustrates the means, reflecting the expected order. Results of an ANOVA (i.e., Analysis of Variance, comparing the four author types) showed that people perceived that writing shares differ between the author types ($F(3,239) = 89.24$, $p < 0.001$). Post-hoc tests (Bonferroni) showed that all author groups were perceived significantly different from each other (each $p < 0.001$) – except for one. The difference between sole AI authorship and AI controlled by human were not different ($p = 0.13$).

Moreover, to evaluate the effect of authorship types on perceptions of human control over AI, respondents had to indicate “who had the final responsibility for the text”, ranging from 1 = AI to 9 = Human) (see Figure 2). For the ANOVA test, the homogeneity of variances was not given (Levene's F= 11.14, $p < 0.001$). To adequately control for this, we used the recommended Welch test and Games-Howell post-hoc tests (Tomarken & Serlin, 1986). Results revealed significant differences between the groups (F_{Welch} (3,131.41) = 13.80, $p < 0.001$). Human control was highest in the case of sole human authorship as no AI was involved, followed by the AI author controlled by a human and the human author supported by AI. Obviously, the lowest level of human control was assigned for sole AI authorship. Post-hoc tests (Games-Howell) showed that human control over AI was significantly higher for human authorship vs. human author supported by AI or AI (each $p < 0.001$), but not significantly different than AI author controlled by a human ($p = 0.25$).

Results

To test H1 and H2 in one comprehensive model, we ran a mediation model (PROCESS model 4 with 5,000 bootstrapped samples and 95% CI's (Hayes, 2018)). The author types were the multicategorical independent variable, message credibility was the mediator, attitude towards the company was the outcome variable, and age, gender, education and country of origin were covariates. Related to the author types, the human author was selected as base case to meet the perceptions and attitudes that were given before AI integration. Compared to a human authored message, respondents perceived an AI author ($b = -0.45$, $p < 0.05$) and a human author supported by AI ($b = -0.71$, $p < 0.005$) as significantly less credible. In contrast, an AI author controlled by human was not perceived significantly different ($p = 0.21$). All covariates had no significant impact on message credibility (each $p > 0.1$). Thus, H1a and H1b could be supported.

In turn, message credibility had a significant impact on attitude towards the company ($b = 0.70$, $p < 0.001$) – supporting H2.

The total effects of authorship types on attitude towards the company were significantly negative for AI authorship ($b = -0.54$, $p < 0.05$) and for the human author supported by AI ($b =$

-0.57, $p < 0.05$), but not significant for an AI author controlled by human ($p = 0.15$). Notably, no direct effects of authorship type on attitude towards the company were significant (each $p > 0.1$), indicating a full mediation for the former two author types. Regarding the covariates, only educational level had a significant impact on message credibility ($b = -0.14$, $p < 0.05$) and no covariate had a total effect on attitude towards the company (each $p > 0.1$).

In sum, both an AI authorship and a human authorship supported by AI have negative effects on readers' attitude towards the company, mediated by lower message credibility perceptions – whereas an AI authorship with human control had not such a negative effect (vs. a human author).

3.2. Study 2

Study 2 aimed to validate the results of Study 1 in another business-related context. In particular, a company's vision statement was chosen as highly relevant message expressing company values and targets. Furthermore, Study 2 assessed the moderating effects of morality of AI use (H3) on message and company evaluations.

Participants and procedure

In exchange for a monetary compensation ($ 0.75), participants from the U.S. were recruited via Amazon mTurk, and randomly assigned to one of the conditions in the 4 (author: human vs. human supported by AI vs. AI controlled by human vs. AI) x 2 (industry: kitchen vs. clothing) between-subjects design. We chose mTurk as one of the most prominent online platform for social science and management research to alter the platform used in study 1. Respondents had to surpass 95% completion rate of former tasks.

After excluding respondents who failed the attention check or the correct recognition of the author(s), the final dataset consisted of n=217 respondents (46.5% females, $M_{age} = 38$ years, SD = 11.37, with an equal or higher than 95% former tasks approval ratio). We altered the industry to control for possible effects due to a more technical or emotional business. Results of two independent samples t-tests showed that the industry type did not influence message credibility ($p = 0.31$), but the message from the fashion industry was rated marginally more positive than from the kitchen industry ($M_{Fashion}$: 5.59, SD: 1.66, $M_{Kitchen}$: 5.26, SD: 1.38, $t(215) = -1.86$, $p < 0.1$).

After accessing the survey, respondents were asked to read a fictitious scenario regarding a company's vision statement that was presented on a website (see Figure 4). We used this context, as it represents a relevant marketing message and a common online content of many companies. Furthermore, a vision statement describes the identity or character of a company and is therefore rather emotionally driven, compared to factual texts such as product specifications, and people might expect a more human touch there. While holding the text equal across the groups, we altered the author types and the industry of the respective company.

Measures

As measures, participants' perceptions about message credibility and attitude towards the company were assessed using the same items as in Study 1. Additionally, perceived morality of companies' AI use to create marketing content was evaluated with a 4-item 7-point semantic differential (Olson et al., 2016). Finally, respondents entered their age, gender and education. All items and factor loadings are shown in Table 2. All psychometric measures were above the recommended levels (see Table 2), suggesting construct reliability and validity (Hulland et al., 2018). Moreover, the experiment groups presented no significant differences regarding the control variables (each $p > 0.1$), suggesting a successful randomization.

[Insert Table 2 about here]

As manipulation check, readers of the different author groups had to evaluate the human (vs. AI) share of input. We used the Welch test and Games-Howell post-hoc tests, because the assumption of homogeneity of variances was violated. The perceived share of human or AI-input differed significantly across the groups (F_{Welch} (3,114.63) = 180.67, $p < 0.001$). Post-hoc tests (Games-Howell) showed that all groups are significantly different from each other ($p < 0.05$). As expected, people in the human author scenario perceived the highest share of human-input (M: 8.29, SD: 1.32), followed by the human author supported by AI (M: 4.42; SD: 2.06) and the "AI author controlled by human" (M: 3.39, SD: 1.88), and perceived the lowest share of human authorship in the AI authorship scenario (M: 2.11; SD: 1.39). Regarding human control over AI (i.e., "who had the final responsibility for the text", ranging from 1 = AI to 9 = Human), results were again significantly different between the groups (F_{Welch} (3,105.34) = 23.92, $p < 0.001$). Human control was highest in the case of sole human authorship as no AI was involved, followed by AI controlled by a human, the human author supported by AI, and was least for sole AI authorship. Post-hoc tests (Games-Howell) showed that human control over AI was significantly higher for human authorship vs. human author supported by AI or vs. AI (each $p < 0.001$), but not significantly different from an AI author controlled by a human ($p = 0.62$).

Scenario realism was assessed with two items from Study 1. Again, all scenarios were perceived as realistic ($\alpha = 0.81$; M: 5.97, SD: 1.00), and realism scores did not differ between the author groups ($p > 0.1$). Respondents confirmed that they "want to know about the use of AI" (M: 5.34, SD: 1.48 on a 7-point scale). Furthermore, the call for transparency (European Parliament, 2023; Jobin et al., 2019) was also reflected, as respondents agreed that "companies should be obliged to disclose the use of AI" (M: 5.18, SD: 1.59). On average, people seem to perceive companies' AI usage as morally rather acceptable (M: 5.00, SD: 1.36), and this perception did not differ among the authorship groups ($p > 0.1$).

Results

To assess the hypothesized effects of the authors on message credibility (H1) and subsequently on attitude towards the company (H2), and the moderating effect of morality (H3) in one comprehensive model, we used a moderated mediation analysis with PROCESS (model 8 with 5,000 bootstrapped samples and 95% CIs (Hayes, 2018)) based on the same setup as in Study 1. As moderator, we included morality of AI use, and we controlled for age, gender and industry type. Table 3 illustrates the results.

Respondents rated the text of sole AI authorship as significantly less credible than a (sole) human-authored text ($b = -2.95$, $p < 0.005$). Again, the collaborative authorships were perceived differently: A text from AI controlled by human was not significantly different from a human authorship ($p = 0.65$), but a text from a human author supported by AI was rated significantly worse ($b = -1.73$, $p < 0.05$). Thus, although people acknowledged that the latter form contains a higher share of human input, this version was rated less credible than a collaboration format with less human input (but human control). Notably, the covariates age, gender, industry type and education had no impact on message credibility ($p > 0.1$).

In turn, message credibility had a significant impact on attitude towards the company ($b = 0.70, p < 0.001$). None of the author types had a direct impact on attitude towards the company (each $p > 0.1$, see Table 3), indicating a full mediation via message credibility. Attitudes towards the company were not influenced by age, gender or education (each $p > 0.1$), while the fashion industry (vs. kitchen) marginally increased the attitudinal evaluations ($b = 0.21$, $p < 0.1$).

In sum, these results support H1 (a and b) and H2 again. Perceptions of human control over AI were found to be more relevant than share of human input when evaluating message credibility. In particular, human-AI collaboration including explicit human control was found to be equally credible as a sole human authorship, whereas the collaboration with higher human input but without such a human control (i.e., human author with AI support) was rated as less credible. Thus, in a collaborative setting, people were found to be rather insensitive to human input, but sensitive to human control over AI (H1). In turn, stronger message credibility led to more favorable attitudes towards the company (H2).

Regarding the hypothesized moderating effects (H3), the interaction of AI authorship (vs. human) x morality was significant ($b = 0.43$, $p < 0.05$), while interactions of the collaborative author types (vs. human) x morality were not significant (each $p > 0.1$). Yet, the conditional indirect effects of the authorships on message credibility offer a more detailed picture. In case of a low (i.e., M-1SD: 3.67) or medium (M: 5.02) perceived morality of AI use, messages from AI and the human authorship supported by AI were perceived as less credible (low morality: b_{AI}: -1.38, $p < 0.001$, $b_{\text{Human supported by AI}}$: -0.92, $p < 0.005$; medium morality: b_{AI}: -0.81, $p < 0.001$, $b_{\text{Human supported by AI}}$: -0.62, $p < 0.005$). Yet, in case of a high perceived morality (i.e., M+1SD: 6.37), these negative effects on message credibility diminished (high morality: $p_{\text{AI}} = 0.53$, $p_{\text{Human supported by AI}} = 0.30$). In contrast, messages from AI controlled by human did not lead to lower message credibility irrespective of the level of perceived morality (each $p > 0.1$). Thus, even for individuals with lower perceptions of morality, the use of AI is not leading to lower credibility perceptions – as long as AI is controlled by a human.

Similarly, indirect effects of the author types on attitude towards the company via message credibility were significantly negative for the AI authorship and the human author supported by AI in case of low and medium morality perceptions; and not significant in case of high morality perceptions (see Table 3). These indirect effects were all insignificant for the AI author controlled by human. Despite this clear pattern, the index of moderated mediation was only significant for the AI authorship (index = 0.30 [0.07; 0.53]). Finally, it should be noted, that morality had an impact on message credibility ($b = 0.40$, $p < 0.005$), but it had no direct effect on people's attitude towards the company ($p = 0.80$). In sum, H3 could be supported for the AI authorship as less credible author type. More generally, perceiving a company's AI use as immoral leads to a stronger credibility devaluation of the author types which lack human control (i.e., AI and Human supported by AI) – whereas this was not the case when AI authorship is controlled by a human. Furthermore, perceiving a company' AI use as morally acceptable diminishes this effect and consumers accept all author forms of authorships as credible authors (see Figure 5).

4. Discussion

Due to rapid technological developments, AI is significantly changing business models and disrupts numerous systems and processes (Daza & Ilozumba, 2022; Fügener et al., 2022; Osburg et al., 2022). However, as people were found to distrust and devalue AI compared to humans, solutions for an ethical and yet image-protecting use and declaration of AI as content author are urgently needed.

Therefore, this study first aimed to examine whether human-AI collaboration could be used to escape people's algorithm aversion. In addition, as scholars acknowledged different forms of collaboration (Bailer et al., 2022; Huang & Rust, 2022), we secondly aimed to investigate the effectiveness of two distinct forms. As specific use case, this study examines effects of AI use as content-generating tool, as these applications (such as ChatGPT) have received considerable attention in business and research to enhance marketing automation (Bailer et al., 2022; Luo et al., 2019; Puntoni et al., 2021). Across two studies with different contexts, we demonstrate that the use and declaration of hybrid authorships could be a solution for the dilemma of how to integrate AI transparently without suffering from negative consumer responses. Thus, in response to our first research aim, human-AI collaborations help exploit the potential for efficiency gains while adhering to the upcoming legislative requirements and circumventing consumers' algorithm aversion.

However, not every form of human-AI collaboration was proven to be effective. In particular, using AI as author with a (final) human control led to comparable message credibility to sole human authorship. In contrast, a human author and AI support led to lower message credibility and reduced consumers' attitudes towards the company – albeit users acknowledged the significantly higher proportion of human input. Thus, people were found to care less about the amount of human input, as long as a human had control over AI. Therefore, this study provides a clear recommendation of how to manage human-AI collaboration (and its declaration), addressing our second research aim.

Moreover, the topic of AI use in consumer-facing business applications is also asserted to activate consumers' evaluations of ethical and moral behavior of companies (Cremer & Kasparov, 2021; Siau & Wang, 2020). Therefore, this study evaluates whether consumers' perceptions of morality of a companies' AI use influence their judgment of different author types on message credibility and their attitudes towards the company. Results show that consumers indeed have different levels of moral acceptance of a company's AI use. In turn, these moral judgments lead to an acceptance or rejection of AI as (co-)author. In particular, when individuals view the use of AI as immoral (i.e. low and medium levels of morality), messages from an AI author or a human author supported by AI are perceived as less credible. However, a message from an AI author controlled by a human does not decrease credibility, regardless of morality perceptions. When consumers find it morally acceptable that companies use AI for content creation (i.e., high levels of morality), the negative effects of AI use or any collaborative form compared to sole human authorship diminish. Thus, morality perceptions play a substantial role when examining effects of AI use in business applications.

Theoretical Implications

This research offers several relevant theoretical contributions. First, this study extends the emerging literature around human-AI collaboration and AI augmentation (Hassani et al., 2020; Huang & Rust, 2022; Zhou et al., 2021) by investigating effects of different human-AI collaboration forms in comparison to human authorship and human replacement (i.e., full AI implementation). More precisely, this paper adds empirical evidence to the sparse literature around human-AI collaboration for business and marketing (Huang & Rust, 2022; Zhou et al., 2021). It also enlarges insights for the related field of journalism and news agencies (Waddell, 2019; Wölker & Powell, 2018), which increasingly rely on AI (Yerushalmy, 2023). Using the pervasive case of content-generating AI (Chui et al., 2022; Dwivedi et al., 2023; Olson, 2022), we also merge this field with business and marketing-related consumer responses.

Related to our second research aim, to the best of our knowledge, this study is the first to compare the effectiveness of different human-AI collaboration forms. Scholars have proposed two main forms of collaboration (Bailer et al., 2022), but research has neglected to evaluate the impact of these forms (e.g., on consumer responses) so far.

Furthermore, the results support the ongoing debate whether AI should augment or replace humans (Hassani et al., 2020; Huang & Rust, 2022; M. Langer & Landers, 2021; Shneiderman, 2020) by showing that AI could take over the task but human control is desired and its replacement would lead to negative consumer responses. Results reveal that people use author labels as evaluation cues to assess the credibility of a message. Thereby, the cue of human control over AI was found to be more effective regarding message credibility and company image than the cue of human input share. Taken together, these results answer scholarly calls how companies should best “distribute work between humans and AI” (Fügener et al., 2022, p. 679) and how “managers can optimize their AI-human intelligence joint workforce” (Huang & Rust, 2022, p. 221). Using the example of text-generating AI, this study also provides insights about how consumers perceive human-AI collaboration related to content creation (as requested by Wölker and Powell, 2018). This study also widens the view on AI automation, because the results overcome the one-dimensional view of recent decades that high AI automation is automatically associated with lower human control and safety (Shneiderman, 2020). Instead, using AI with human control offers an escape from this postulated trade-off situation.

Second, this study adds to the literature around individuals' responses to AI. In line with related studies, our results support the phenomenon of algorithm aversion (Burton et al., 2020; Dietvorst et al., 2015; Longoni et al., 2019; Luo et al., 2019), leading to a negative impact of AI (vs. human) authorship on message credibility and company evaluations. However, although consumers were found to have an algorithm aversion, they cared less about the actual amount of human or AI input, but based their evaluation of message credibility on the level of (human) control over AI. Our findings support the results of Dietvorst et al. (2016) that people need a feeling of control over AI, and offer an explanation for the results. Notably, compared to Dietvorst et al. (2016), in our setting, the control over AI was delegated to a human author from the company (as source creator), rather than executed by the consumers themselves. In line with the compensatory control theory (Landau et al., 2015), we show that reducing AI aversion can even be achieved via creating perceptions of secondary control. It also supports the notion of scholars that "even an illusion of autonomy will remedy algorithm aversion" (Burton et al., 2020, p. 224). Further, our results support Longoni et al.'s (2019) findings that when AI collaborates with humans (rather than to replace them), negative effects of AI diminish. Particularly, our results suggest that perceived human control over AI is the underlying mechanism of this alleviation. It should be noted that the disclosure of human control over AI allows users to interpret different types of human control (see Nyholm (2022) for an overview), as the form of control implementation or details of its execution are deliberately not presented.

Third, we extend findings to the new field of AI ethics and link it to business and marketing strategy (Cremer & Kasparov, 2021; Hagendorff, 2020; Siau & Wang, 2020). Scholars acknowledge that the implementation of AI in business and consumer-related tasks creates ethical issues for the organization and marketing management (Puntoni et al., 2021; Siau & Wang, 2020). Recently, scholars increasingly investigate how to handle these issues. For instance, Wirtz et al. (2022) discuss that management should support structures and human personnel as governance mechanism to enable corporate digital responsibility. Our results support this notion and underline the demand for a human control function to "provide human oversight of AI and refinement of data capture and technologies" (Wirtz et al., 2022, 9).

Finally, the integration of people's perception of morality of AI use in companies shows that the moral judgment of AI tools in marketing is diverse in society and influences acceptance or rejection of AI and downstream impacts on company evaluations. Therefore, consumers' evaluations about morality of AI use is an important dimension to be considered in AI-related research.

Managerial implications

AI-driven tools offer managers manifold opportunities to raise efficiencies and profitability, but research has also shown that AI use could alienate customers and harm business (Luo et al., 2019; Puntoni et al., 2021). For instance, companies could use generative AI for content creation in various business fields (see Chui et al. (2022) for details). To use the advantages of AI, insights about proper implementation of AI in business strategy and consumer-facing processes are crucial for the company image and customer retention. Today, most companies are users of AI technology, and not creators of it. Due to the high complexity, time and cost efforts, the design of AI needs specialized labor – which means that managers have no or only limited possibilities to build or customize AI tools (Kozinets & Gretzel, 2021). However, managers could decide about whether and how AI and humans should work together in their processes, which makes guidelines for human-AI collaboration highly relevant for them.

First, results of this study suggest that the use of human-AI collaboration is an effective option to use advantages of AI-driven process automation and, at the same time, to protect the company image. In view of the ethical and upcoming legal obligation to disclose AI use, this research offers managerial guidance to optimize teamwork of AI and humans (Huang & Rust, 2022; Rust, 2020). By comparing two main forms of human-AI collaboration (i.e., (1) AI augmentation and (2) AI takeover with human control) (Huang & Rust, 2022; Longoni et al., 2019; Osburg et al., 2022), this study showed that the latter form is more beneficial as it did not harm message credibility perceptions and company image. Moreover, in contrast to full AI use or AI augmentation, AI takeover with human control was also perceived similarly to sole human authorship irrespective of readers' perceptions of morality of AI use. This means, that managers could use almost the full potential of AI automation – as long as they install a final human control as well. This is a win-win situation for managers, as they can be ethical and transparently declarare AI use and at the same time, use AI automation to a high level. Moreover, using a human lead author and AI augmentation was found to have negative impacts on message credibility perceptions and attitudes towards the company. Thus, managers should emphasize the human control function and not the input level. By using a human control function, companies also follow suggestions of scholars to harness higher levels of AI autonomy (Osburg et al., 2022; Santoni de Sio & van den Hoven, 2018). Such integration of a human control function enables companies to assume their digital responsibility (Wirtz et al., 2022).

Second, managers and software designers need to be aware that consumers have different judgments of how morally acceptable their AI use is. Interestingly, respondents in Study 2 rated the use of AI by companies as quite morally acceptable on average. However, the individual moral judgments differed and influenced the evaluation of the credibility of the message and the company image. In order to support the moral acceptance of AI use, managers might integrate a message next to the author description explaining the reasons of AI use. For instance, consumers are supposed to understand and accept the need for AI to generate highly personalized content based on individual preferences and past behavior (Puntoni et al., 2021). Moreover, several scholars assert that using explainable AI (i.e., provide information how AI makes decisions and performs actions) could lead to more favorable consumer reactions such as trust in AI or fairness perceptions (Rai, 2020).

Limitations and further research

This research has some limitations, which direct to interesting future research opportunities.

First, both studies relied on standardized and one-way communication with company-generated content. However, one of the strengths of AI is the ability to build personalized content based on big data and past consumer behavior (Puntoni et al., 2021). Future studies could examine evaluations of AI use for personalized content, and incorporate e.g., people's trade-offs between appreciating more appropriate information and privacy concerns.

Second, this research focused on consumers' credibility assessment and attitudes towards the company as sender of the message. Future research could investigate other outcomes, for instance actual behavior such as adherence to product recommendations or click rates on web links in the message.

Third, as the use and transparent declaration of AI is touching the field of AI ethics, we integrated the moderator 'morality of AI use' (Cremer & Kasparov, 2021; Hagendorff, 2020). Although this variable was found to differentiate consumers' author evaluations, the judgment whether companies' AI use is (un-)ethical could vary drastically depending on context, cultural environment and personality, among others (Zhou et al., 2021). Therefore, future studies might assess the effects of further variations, for instance different settings including morally critical products and services (such as messages related to weapons or politics), varying consumer-company relationships, different countries with divergent ethical norms, or individual-related factors. For instance, consumers' topic involvement might influence their evaluation of the message credibility depending on the author. For highly relevant personal or sensitive topics, people may be less willing to accept AI and maybe even feel devalued being served by a machine. In contrast, for technology-related topics or high-tech companies, people might even admire AI-created content or human-AI collaboration as expressions of an innovative and future-oriented business.

Fourth, in our study, the disclosure of human control over AI in the scenarios does deliberately not include the form of control implementation or details of its execution. However, according to Nyholm (2022), different forms of control exist and might thus be evaluated differently. Future studies could evaluate the impact of different control forms or control framings on consumers' perceptions and company assessments.

Finally, this study uses a cross-sectional design and represents a current snapshot on this dynamic topic. As AI is continuously and rapidly evolving, future research might investigate long-term effects, for instance whether familiarization with AI-generated content leads to more favorable AI evaluations. Parallel to the growth of AI tools, research from different disciplines should orchestrate efforts to explore further effects of human-AI collaborations and the human control function over AI, to achieve an ethical and beneficial use of AI.

5. References

Alfonseca, M., Cebrian, M., Fernandez Anta, A., Coviello, L., Abeliuk, A., & Rahwan, I. (2021). Superintelligence Cannot be Contained: Lessons from Computability Theory. *Journal of Artificial Intelligence Research*, *70*, 65–76. https://doi.org/10.1613/jair.1.12202

Ameen, N., Sharma, G. D., Tarba, S., Rao, A., & Chopra, R. (2022). Toward advancing theory on creativity in marketing and artificial intelligence. *Psychology & Marketing*, *39*(9), 1802–1825. https://doi.org/10.1002/mar.21699

Appelman, A., & Sundar, S. S. (2016). Measuring Message Credibility. *Journalism & Mass Communication Quarterly*, *93*(1), 59–79. https://doi.org/10.1177/1077699015606057

Bailer, W., Thallinger, G., Krawarik, V., Schell, K., & Ertelthalner, V. (2022). AI for the Media Industry: Application Potential and Automation Levels. In B. Þór Jónsson, C. Gurrin, M.-T. Tran, D.-T. Dang-Nguyen, A. M.-C. Hu, B. Huynh Thi Thanh, & B. Huet (Eds.), *Lecture Notes in Computer Science. MultiMedia Modeling* (Vol. 13141, pp. 109–118). Springer International Publishing. https://doi.org/10.1007/978-3-030-98358-1_9

Bentele, G. (1998). Vertrauen / Glaubwürdigkeit. In O. Jarren, U. Sarcinelli, & U. Saxer (Eds.), *Politische Kommunikation in der demokratischen Gesellschaft: Ein Handbuch mit Lexikonteil* (1st ed., pp. 305–311). Westdt. Verl. https://link.springer.com/book/10.1007%2F978-3-322-80348-1?page=2#toc

Billore, S., & Anisimova, T. (2021). Panic buying research: A systematic literature review and future research agenda. *International Journal of Consumer Studies*, *45*(4), 777–804. https://doi.org/10.1111/ijcs.12669

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

California Legislative, (2018 & rev. SB-1001). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Chan-Olmsted, S. M. (2019). A Review of Artificial Intelligence Adoptions in the Media Industry. *International Journal on Media Management*, *21*(3-4), 193–215. https://doi.org/10.1080/14241277.2019.1695619

Chen, C. Y., Lee, L., & Yap, A. J. (2017). Control Deprivation Motivates Acquisition of Utilitarian Products. *Journal of Consumer Research*, *43*, 1031-1047. https://doi.org/10.1093/jcr/ucw068

Chui, M., Roberts, R., & Yee, L. (2022). *Generative AI is Here: How Tools Like ChatGpt Could Change Your Business*. www.mckinsey.com/capabilities/quantumblack/our-insights/generative-ai-is-here-how-tools-like -chatgpt-could-change-your-business/

Cremer, D. de, & Kasparov, G. (2021). The ethical AI—paradox: why better technology needs more and not less human responsibility. *AI and Ethics.* Advance online publication. https://doi.org/10.1007/s43681-021-00075-y

Creyer, E. H., & Ross, W. T. (1997). The influence of firm behavior on purchase intention: do consumers really care about business ethics? *Journal of Consumer Marketing*, *14*(6), 421–432. https://doi.org/10.1108/07363769710185999

Cutright, K. M., & Wu, E. C. (2023). In and out of control: Personal control and consumer behavior. *Consumer Psychology Review*, *6*(1), 33–51. https://doi.org/10.1002/arcp.1083

Darke, P. R., Ashworth, L., & Ritchie, R. J. (2008). Damage from Corrective Advertising: Causes and Cures. *Journal of Marketing*, *72*(6), 81–97.

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, *48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

Daza, M. T., & Ilozumba, U. J. (2022). A survey of AI ethics in business literature: Maps and trends between 2000 and 2021. *Frontiers in Psychology*, *13*, 1042661. https://doi.org/10.3389/fpsyg.2022.1042661

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology. General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

European Parliament. (2023). *EU AI Act: first regulation on artificial intelligence.* European Parliament. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Fan, M., Yang, X., Yu, T., Liao, Q. V., & Zhao, J. (2022). Human-AI Collaboration for UX Evaluation: Effects of Explanation and Synchronization. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–32. https://doi.org/10.1145/3512943

Finkel, M., & Krämer, N. C. (2022). Humanoid Robots – Artificial. Human-like. Credible? Empirical Comparisons of Source Credibility Attributions Between Humans, Humanoid Robots, and Non-human-like Devices. *International Journal of Social Robotics*, *14*(6), 1397–1411. https://doi.org/10.1007/s12369-022-00879-w

Fisher, J. (2004). Social Responsibility and Ethics: Clarifying the Concepts. *Journal of Business Ethics*, *52*(4), 381–390. https://doi.org/10.1007/s10551-004-2545-y

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Friesen, J. P., Kay, A. C., Eibach, R. P., & Galinsky, A. D. (2014). Seeking structure in social organization: Compensatory control and the psychological advantages of hierarchy. *Journal of Personality and Social Psychology*, *106*(4), 590–609. https://doi.org/10.1037/a0035620

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*, *33*(2), 678–696. https://doi.org/10.1287/isre.2021.1079

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*, 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Graefe, A., & Bohlken, N. (2020). Automated Journalism: A Meta-Analysis of Readers' Perceptions of Human-Written in Comparison to Automated News. *Media and Communication*, *8*(3), 50–59. https://doi.org/10.17645/mac.v8i3.3019

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2016). Readers′ perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, *19*(5), 595–610. https://doi.org/10.1177/1464884916641269

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Harmon, R. R., & Coney, K. A. (1982). The Persuasive Effects of Source Credibility in Buy and Lease Situations. *Journal of Marketing Research*, *19*(2), 255–260. https://doi.org/10.1177/002224378201900209

Hassani, H., Silva, E. S., Unger, S., TajMazinani, M., & Mac Feely, S. (2020). Artificial Intelligence (AI) or Intelligence Augmentation (IA): What Is the Future? *AI*, *1*(2), 143–155. https://doi.org/10.3390/ai1020008

Hayes, A. F. (2018). *Introduction to Mediation, Moderation and Conditional Process Analysis. A Regression-Based Approach* (2nd edition). The Guilford Press.

Hermann, E. (2022). Leveraging Artificial Intelligence in Marketing for Social Good-An Ethical Perspective. *Journal of Business Ethics : JBE*, *179*(1), 43–61. https://doi.org/10.1007/s10551-021-04843-y

Hovland, C. I., Janis, I. L., & Kelley, H. L. (1953). *Communication and persuasion; psychological studies of opinion change.* Yale University Press.

Huang, M.-H., & Rust, R. T. (2022). A Framework for Collaborative Artificial Intelligence in Marketing. *Journal of Retailing*, *98*(2), 209–223. https://doi.org/10.1016/j.jretai.2021.03.001

Hulland, J., Baumgartner, H., & Smith, K. M. (2018). Marketing survey research best practices: evidence and recommendations from a review of JAMS articles. *Journal of the Academy of Marketing Science*, *46*(1), 92–108. https://doi.org/10.1007/s11747-017-0532-y

Illia, L., Colleoni, E., & Zyglidopoulos, S. (2023). Ethical implications of text generation in the age of artificial intelligence. *Business Ethics, the Environment & Responsibility*, *32*(1), 201–210. https://doi.org/10.1111/beer.12479

Ismagilova, E., Slade, E., Rana, N. P., & Dwivedi, Y. K. (2020). The effect of characteristics of source credibility on consumer behaviour: A meta-analysis. *Journal of Retailing and Consumer Services*, *53*, 101736. https://doi.org/10.1016/j.jretconser.2019.01.005

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kahnt, I. (2019). Künstliche Intelligenz im Content Marketing. In M. Wesselmann (Ed.), *Content gekonnt: Strategie, Organisation, Umsetzung, ROI-Messung und Fallbeispiele aus der Praxis* (pp. 211–225). Springer Gabler.

Kay, A. C., Gaucher, D., Napier, J. L., Callan, M. J., & Laurin, K. (2008). God and the government: Testing a compensatory control mechanism for the support of external systems. *Journal of Personality and Social Psychology*, *95*(1), 18–35. https://doi.org/10.1037/0022-3514.95.1.18

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, *114*, 106553. https://doi.org/10.1016/j.chb.2020.106553

Kozinets, R. V., & Gretzel, U. (2021). Commentary: Artificial Intelligence: The Marketer's Dilemma. *Journal of Marketing*, *85*(1), 156–159. https://doi.org/10.1177/0022242920972933

Lai, Y., Kankanhalli, A., & Ong, D. (2021). Human-AI Collaboration in Healthcare: A Review and Research Agenda. In *Proceedings of the Annual Hawaii International Conference on System Sciences* (pp. 390–399). Hawaii International Conference on System Sciences. https://doi.org/10.24251/hicss.2021.046

Landau, M. J., Kay, A. C., & Whitson, J. A. (2015). Compensatory control and the appeal of a structured world. *Psychological Bulletin*, *141*(3), 694–722. https://doi.org/10.1037/a0038703

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32*(2), 311–328. https://doi.org/10.1037/0022-3514.32.2.311

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior*, *123*, 106878. https://doi.org/10.1016/j.chb.2021.106878

Larivière, B., Bowen, D., Andreassen, T. W., Kunz, W., Sirianni, N. J., Voss, C., Wünderlich, N. V., & Keyser, A. de (2017). "Service Encounter 2.0": An investigation into the roles of technology, employees and customers. *Journal of Business Research*, *79*, 238–246. https://doi.org/10.1016/j.jbusres.2017.03.008

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2). https://doi.org/10.14763/2020.2.1469

Libai, B., Bart, Y., Gensler, S., Hofacker, C. F., Kaplan, A., Kötterheinrich, K., & Kroll, E. B. (2020). Brave New World? On AI and the Management of Customer Relationships. *Journal of Interactive Marketing*, *51*, 44–56. https://doi.org/10.1016/j.intmar.2020.04.002

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Longoni, C., & Cian, L. (2022). Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The “Word-of-Machine” Effect. *Journal of Marketing*, *86*(1), 91–108. https://doi.org/10.1177/0022242920957347

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science.* Advance online publication. https://doi.org/10.1287/mksc.2019.1192

Mellish, C., Scott, D., Cahill, L., Paiva, D., Evans, R., & Reape, M. (2006). A Reference Architecture for Natural Language Generation Systems. *Natural Language Engineering*, *12*(1), 1–34.

Nelson, M. R., & Park, J. (2015). Publicity as Covert Marketing? The Role of Persuasion Knowledge and Ethical Perceptions on Beliefs and Credibility in a Video News Release Story. *Journal of Business Ethics*, *130*(2), 327–341. https://doi.org/10.1007/s10551-014-2227-3

Newman, A., & Jones, R. (2006). Authorship of research papers: Ethical and professional issues for short-term researchers. *Journal of Medical Ethics*, *32*(7), 420–423. https://doi.org/10.1136/jme.2005.012757

Nyholm, S. (2022). A new control problem? Humanoid robots, artificial intelligence, and the value of control. *AI and Ethics.* Advance online publication. https://doi.org/10.1007/s43681-022-00231-y

Ohanian, R. (1990). Construction and Validation of a Scale to Measure Celebrity Endorsers’ Perceived Expertise, Trustworthiness, and Attractiveness. *Journal of Advertising*, *19*(3), 39–52. https://doi.org/10.1080/00913367.1990.10673191

Olson, P. (2022). *Google Faces a Serious Threat From ChatGPT.* https://www.washingtonpost.com/business/energy/google-faces-a-serious-threat-from-chatgpt/2022/12/07/363d2440-75f5-11ed-a199-927b334b939f_story.html

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409. https://doi.org/10.1002/bdm.637

Osburg, V.-S., Yoganathan, V., Kunz, W. H., & Tarba, S. (2022). Can (A)I Give You a Ride? Development and Validation of the CRUISE Framework for Autonomous Vehicle Services. *Journal of Service Research*, 109467052211182. https://doi.org/10.1177/10946705221118233

Parasuraman, A., & Colby, C. L. (2015). An Updated and Streamlined Technology Readiness Index. *Journal of Service Research*, *18*(1), 59–74. https://doi.org/10.1177/1094670514539730

Paschen, J., Wilson, M., & Ferreira, J. J. (2020). Collaborative intelligence: How human and artificial intelligence create value along the B2B sales funnel. *Business Horizons*, *63*(3), 403–414. https://doi.org/10.1016/j.bushor.2020.01.003

Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and Artificial Intelligence: An Experiential Perspective. *Journal of Marketing*, *85*(1), 131–151. https://doi.org/10.1177/0022242920953847

Raftopoulos, Marigo, & Hamari, J. (2023). Human-AI collaboration in organisations: A literature review on enabling value creation. *European Conference on Information Systems*, *31*.

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rust, R. T. (2020). The future of marketing. *International Journal of Research in Marketing*, *37*(1), 15–26. https://doi.org/10.1016/j.ijresmar.2019.08.002

Schermerhorn, J. R. (2002). *Management* (7th ed.). Wiley.

Schleith, J., Hoffmann, H., Norkute, M., & Cechmanek, B. (2022). *Human in the loop information extraction increases efficiency and trust.* https://doi.org/10.18420/muc2022-mci-ws12-249

Seo, S. H., Griffin, K., Young, J. E., Bunt, A., Prentice, S., & Loureiro-Rodríguez, V. (2018). Investigating People's Rapport Building and Hindering Behaviors When Working with a Collaborative Robot. *International Journal of Social Robotics*, *10*(1), 147–161. https://doi.org/10.1007/s12369-017-0441-8

Shaw, W., & Barry, V. (2014). *Moral issues in business* (13th ed.). Cengage Learning, Inc.

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics. *Journal of Database Management*, *31*(2), 74–87. https://doi.org/10.4018/JDM.2020040105

Skinner, E. A. (1996). A guide to constructs of control. *Journal of Personality and Social Psychology*, *71*(3), 549–570. https://doi.org/10.1037/0022-3514.71.3.549

Sowa, K., Przegalinska, A., & Ciechanowski, L. (2021). Cobots in knowledge work. *Journal of Business Research*, *125*, 135–142. https://doi.org/10.1016/j.jbusres.2020.11.038

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

Tobia, K., Nielsen, A., & Stremitzer, A. (2021). When Does Physician Use of AI Increase Liability? *Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine*, *62*(1), 17–21. https://doi.org/10.2967/jnumed.120.256032

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*(1), 90–99. https://doi.org/10.1037/0033-2909.99.1.90

Vakratsas, D., & Wang, X. (2021). Artificial Intelligence in Advertising Creativity. *Journal of Advertising*, *50*(1), 39–51. https://doi.org/10.1080/00913367.2020.1843090

Visentin, M., Pizzi, G., & Pichierri, M. (2019). Fake News, Real Problems for Brands: The Impact of Content Truthfulness and Source Credibility on consumers' Behavioral Intentions toward the Advertised Brands. *Journal of Interactive Marketing*, *45*, 99–112. https://doi.org/10.1016/j.intmar.2018.09.001

Waddell, T. F. (2018). A Robot Wrote This? *Digital Journalism*, *6*(2), 236–255. https://doi.org/10.1080/21670811.2017.1384319

Waddell, T. F. (2019). Can an Algorithm Reduce the Perceived Bias of News? Testing the Effect of Machine Attribution on News Readers' Evaluations of Bias, Anthropomorphism, and Credibility. *Journalism & Mass Communication Quarterly*, *96*(1), 82–100. https://doi.org/10.1177/1077699018815891

Wagner, T., Hennig-Thurau, T., & Rudolph, T. (2009). Does Customer Demotion Jeopardize Loyalty? *Journal of Marketing*, *73*(3), 69–85. https://doi.org/10.1509/jmkg.73.3.69

Wang, A. Y., Wang, D., Drozdal, J., Muller, M., Park, S., Weisz, J. D., Liu, X., Wu, L., & Dugan, C. (2022). Documentation Matters: Human-Centered AI System to Assist Data Science Code Documentation in Computational Notebooks. *ACM Transactions on Computer-Human Interaction*, *29*(2), 1–33. https://doi.org/10.1145/3489465

West, M. D. (1994). Validating a Scale for the Measurement of Credibility: A Covariance Structure Modeling Approach. *Journalism Quarterly*, *71*(1), 159–168. https://doi.org/10.1177/107769909407100115

Wien, A. H., & Peluso, A. M. (2021). Influence of human versus AI recommenders: The roles of product type and cognitive processes. *Journal of Business Research*, *137*, 13–27. https://doi.org/10.1016/j.jbusres.2021.08.016

Wirtz, J., Kunz, W. H., Hartley, N., & Tarbit, J. (2022). Corporate Digital Responsibility in Service Firms and Their Ecosystems. *Journal of Service Research*, 109467052211304. https://doi.org/10.1177/10946705221130467

Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., Lu, V. N., Paluch, S., & Martins, A. (2018). Brave new world: service robots in the frontline. *Journal of Service Management*, *29*(5), 907–931. https://doi.org/10.1108/JOSM-04-2018-0119

Wölker, A., & Powell, T. E. (2018). Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*, 146488491875707. https://doi.org/10.1177/1464884918757072

Wu, Y. (2019). Is Automated Journalistic Writing Less Biased? An Experimental Test of Auto-Written and Human-Written News Stories. *Journalism Practice*, *14*(8), 1008–1028. https://doi.org/10.1080/17512786.2019.1682940

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414. https://doi.org/10.1002/bdm.2118

Yerushalmy, J. (2023). *German publisher Axel Springer says Journalists could be replaced by AI*. https://www.theguardian.com/technology/2023/mar/01/german-publisher-axel-springer-says-journalists-could-be-replaced-by-ai

Zanzotto, F. M. (2019). Viewpoint: Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research*, *64*, 243–252. https://doi.org/10.1613/jair.1.11345

Zhou, L., Paul, S., Demirkan, H., Yuan, L., Spohrer, J., Zhou, M., & Basu, J. (2021). Intelligence Augmentation: Towards Building Human- machine Symbiotic Relationship. *AIS Transactions on Human-Computer Interaction*, *13*(2), 243–264. https://doi.org/10.17705/1thci.00149

Figure 1. Conceptual model

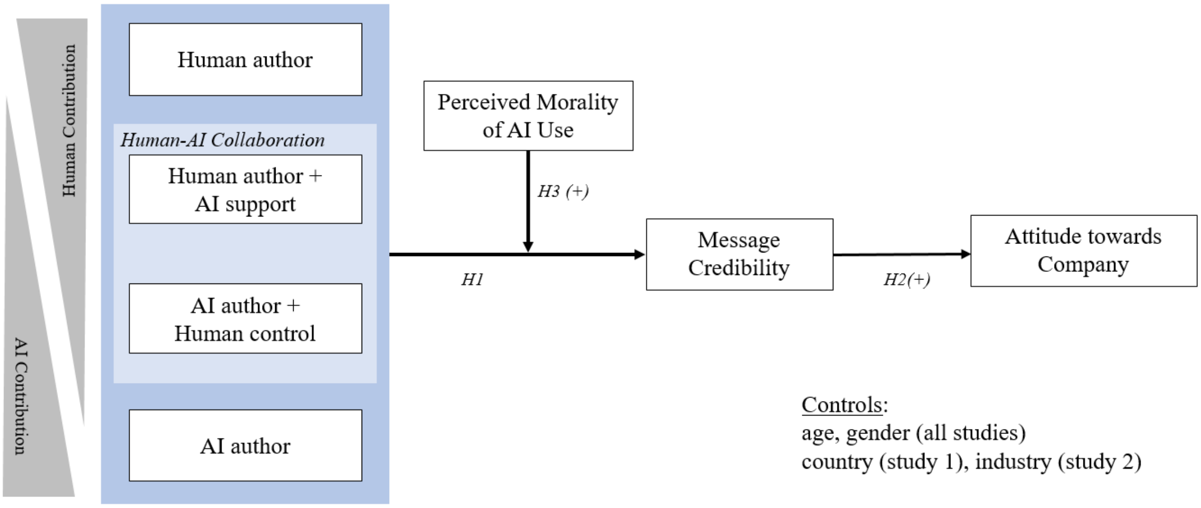
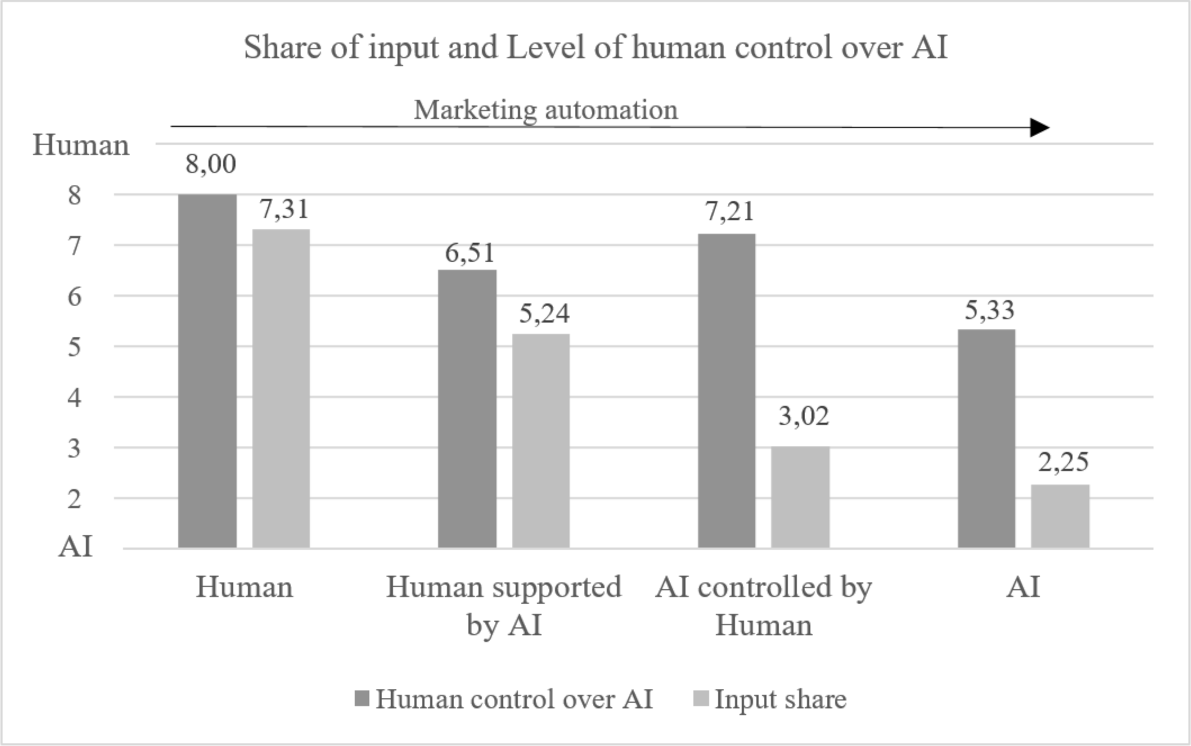
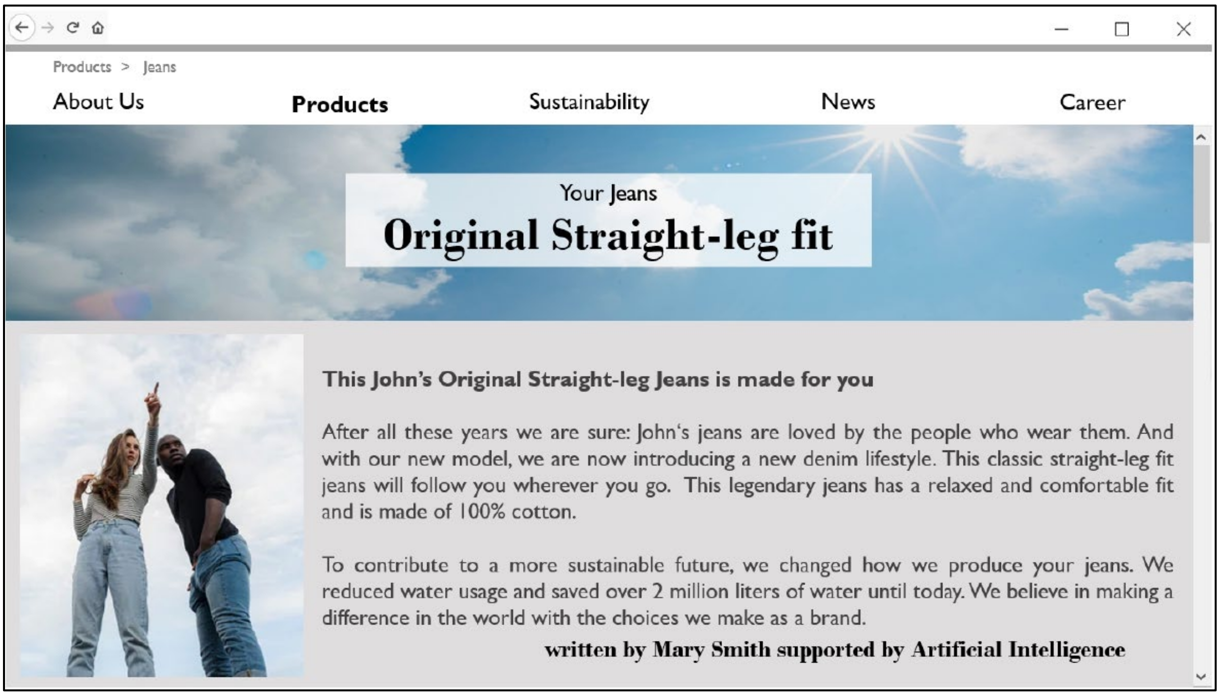


Figure 2. Study 1. Readers' perceptions of Share of Input and Level of Control.



Note. Scale ranging from 1= AI to 9 = Human.

Figure 3. Study 1. Exemplary Scenario.



Your Jeans

Original Straight-leg fit

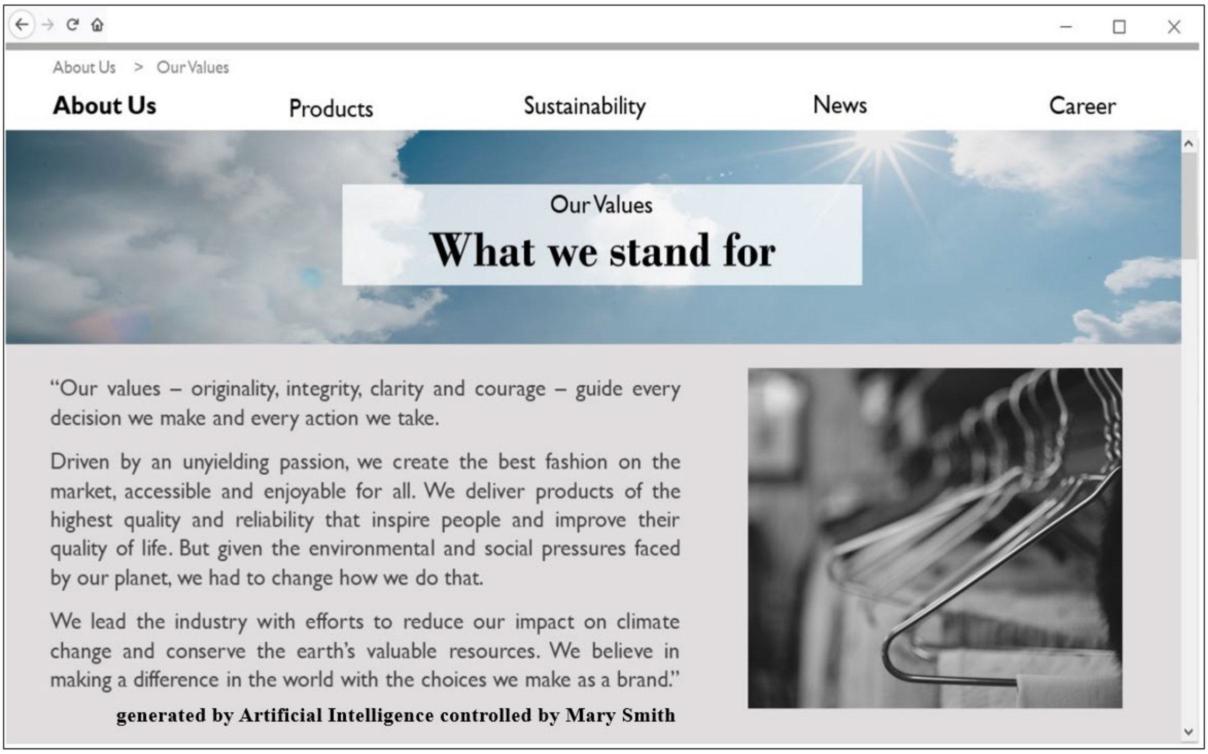
This John's Original Straight-leg Jeans is made for you

After all these years we are sure: John's jeans are loved by the people who wear them. And with our new model, we are now introducing a new denim lifestyle. This classic straight-leg fit jeans will follow you wherever you go. This legendary jeans has a relaxed and comfortable fit and is made of 100% cotton.

To contribute to a more sustainable future, we changed how we produce your jeans. We reduced water usage and saved over 2 million liters of water until today. We believe in making a difference in the world with the choices we make as a brand.

written by Mary Smith supported by Artificial Intelligence

Figure 4. Study 2. Exemplary Scenario.



Our Values

What we stand for

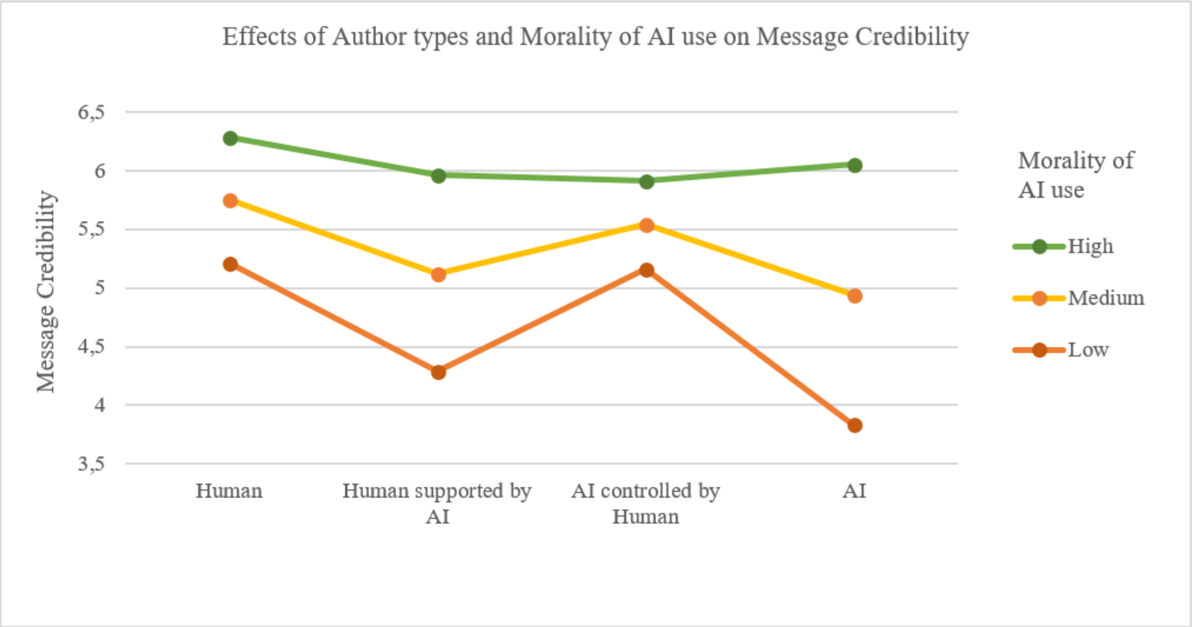
"Our values – originality, integrity, clarity and courage – guide every decision we make and every action we take.

Driven by an unyielding passion, we create the best fashion on the market, accessible and enjoyable for all. We deliver products of the highest quality and reliability that inspire people and improve their quality of life. But given the environmental and social pressures faced by our planet, we had to change how we do that.

We lead the industry with efforts to reduce our impact on climate change and conserve the earth's valuable resources. We believe in making a difference in the world with the choices we make as a brand."

generated by Artificial Intelligence controlled by Mary Smith

Figure 5. Study 2. Influence of Morality of AI use on relationship between authors and message credibility



Note. Values represent estimated marginal means, with age, gender and industry as covariates. Morality of AI levels are: low (M-1SD: 3.67), medium (M: 5.02), and high (M+1SD: 6.37).

Table 1. Selected literature related to Human-AI collaboration

Research Streams	Source	Topic	Literature Field(s)	Mediator(s), Moderator(s)	Main Findings
Conceptual	Huang and Rust (2022)	Conceptual framework for collaborative AI in marketing	Marketing	-	AI advances from mechanical, to thinking, to feeling intelligence. Human-AI collaboration can be achieved through (1) using the respective strengths of human or AI, (2) using lower-level AI to augment higher-level human intelligence, or (3) using AI to automate lower intelligence processes and humans focus on higher intelligence tasks. Possible boundary conditions such as task (un-)desirability are discussed.
	Davenport et al., 2019	Multidimensional framework which integrates AI intelligence levels, task types and appearance	Marketing	-	AI will influence marketing strategies and consumer behavior in manifold ways. Authors suggest that human-AI collaboration is more effective than human replacement, and ethical issues need to be considered cautiously.
	Paschen et al., 2020	Conceptual framework for human-AI collaboration based on B2B sales funnel	Marketing, Sales	-	AI could add significant value in every stage of the sales funnel. Specific contributions of AI and humans in the collaboration are outlined, and benefits for business are derived.
	Raftopoulos and Hamari, 2023	Business potentials for value creation due to human-AI collaboration	IS, general management	-	Human-AI collaboration offers substantial potential to create sustainable business value. Authors posit four main enablers for successful value creation: strategic positioning, human engagement, organizational evolution and technology development (including user acceptance to overcome AI aversion or AI anxiety).
	Sundar, 2020	Framework regarding the psychology of Human–AI interaction	Media, HCI	-	The article proposes a framework to investigate synergies of human agency and machine agency. It presents symbolic and enabling effects of AI-driven media affordances on user perceptions, experiences and engagement with AI medium. The author suggests that human-AI co-creation of media will significantly shape these mentioned outcomes.
	Zanzotto et al., 2019	Human-in-the-loop paradigm as collaborative format in the economy	AI research	-	AI has disruptive potential for economy and labor market due to self-learning characteristics, human-AI collaboration as proposed system helps human knowledge creators to participate in AI-driven revenues and wealth.

Table 1 (continued). Selected literature related to Human-AI collaboration

	Zhou et al., 2021	Conceptual framework for symbiotic human-AI collaboration	IT, HCI	-	Intelligence Augmentation (IA), meaning that AI is used to amplify human abilities is outlined as recommended and highly likely format of human-AI collaboration. IA is expected to increase business' competitiveness and value.
Empirical	Fan et al., 2022	AI tools as support for user experience evaluations	IT, HCI	-	AI-assisted user experience (UX) assessments increased efficiency. AI was particularly more supportive when the AI work is presented with explanations (i.e., how AI identified UX problems).
	Fügener et al., 2021	Productive delegation of tasks in human-AI collaboration	IS, HCI, general management	-	Human-AI collaboration could outperform AI, which outperforms humans in classification tasks. However, combined performance is only improved when AI delegates work to humans (and not vice versa), because humans could not assess their own capabilities correctly.
	Lai et al., 2021	Literature review on human-AI collaboration in healthcare	Healthcare, Medical	-	Literature on healthcare-related collaboration discusses different AI areas and technologies, focus on specific diseases, different outcomes (e.g., task performance, completion time, learning, usability, acceptance) and stakeholders involved (e.g., healthcare professionals, patients, clinical researchers).
	Longoni et al., 2019	Consumer responses to human-AI collaborative medical advice	Healthcare, medical	Uniqueness neglect,	Human-AI collaboration in giving medical advice eliminates individuals' resistance to accept AI-advice. Thus, AI acceptance is given when it supports (vs. replaces) a human doctor.
	Schleith et al., 2022	Human-AI collaboration regarding information extraction systems	Information systems	-	Installing specific forms of human-AI collaboration (i.e., manual-rule-based-, black-box-review) offers users more understanding and perceived control, which leads to higher trust and joint task efficiency.
	Sowa et al., 2021	Productivity due to human-AI collaboration in managerial tasks	Management	-	Two studies confirm that human-AI collaboration increases perceived productivity. Moreover, people are generally ready to use AI in managerial professions and that collaborative tools should be personalized. Higher sense of agency and ability to influence tool design is suggested to improve the collaboration.
	Waddell, 2019	Human-AI collaboration effects on readers' article credibility evaluations	Journalism	Med.: Perceived bias, Anthropomorphism	Author find two mediation paths: human-AI collaboration for news creation was perceived as less biased (vs. human author), leading to positive effects on article (message) credibility. However, collaborative format was also perceived as less anthropomorphic than a human author,

Table 1 (continued). Selected literature related to Human-AI collaboration

				creating negative effects on message credibility. Topic type or story context were not significantly influencing these effects.
Wang et al., 2022	Human-AI collaboration in code documentation as part of data science	Data Science	-	AI-assisted coding and code documentation supports human researcher and leads to improved quality and productivity. In particular, human-in-the-loop (and similar) design principles support these goals.
Wölker and Powell, 2018	Perceived article credibility and news selection of automated journalism	Journalism	Med: message credibility, source credibility, Mod.: Article type	Credibility perceptions were not significantly different between human, AI, or human-AI collaborative authorship formats. Moreover, news selection is not influenced by message or source credibility.
Our Study	Consumer perceptions of two forms of Human-AI collaborations on firm evaluation	Marketing, Management	Med.: Message credibility, Mod.: perceived morality of companies' AI use	Human-AI collaboration can alleviate negative consumer responses regarding AI use, but only when the collaboration indicates human control over AI (vs. AI supporting a human). When people perceive a high morality towards companies' AI use, negative effects of any author type diminish.

Note. Mod. = Moderator(s), Med. = Mediator(s), IS = information systems, IT = information technology, HCI = human-computer interaction

Table 2. Scale items and statistics

Construct name and items	Standardized loadings	
	Study 1	Study 2
Message Credibility (Study 1 / Study 2: α = .88/91; CR = .88/.86; AVE = .65/.61)		
This text …		
• … is generally truthful.	.75	.75
• … leaves one feeling accurately informed.	.79	.74
• … is believable.	.84	.85
• … is authentic.	.83	.77
Attitude towards the company (Study 1 / Study 2: α = .95/.89; CR = .92/.85; AVE = .79/.66)		
• This company is a good company	.88	.85
• This company is a nice company.	.90	.84
• I like the company	.89	.74
Morality of AI use (Study 2: α = .93; CR = .93; AVE = .77)		
Companies using artificial intelligence (AI) in marketing texts are...		
• Cruel (1) vs. Kind-hearted (7)		.88
• Immoral (1) vs. Moral (7)		.90
• Uncaring (1) vs. Caring (7)		.83
• Unethical (1) vs. Ethical (7)		.89

Table 3. Study 2. Conditional process model for message credibility as mediator, morality of AI use as moderator, and attitude towards the company as outcome

	Mediator Message credibility		Outcome Attitude towards the company	
	b	*t*	*b*	*t*
X1: Human supported by AI vs. Human	-1.72	-2.07*	-.48	-.71 $^{\text{n.s.}}$
X2: AI controlled by Human vs. Human	.39	.47 $^{\text{n.s.}}$	-.71	-1.07 $^{\text{n.s.}}$
X3: AI vs. Human	-2.93	-3.33**	-.01	-.01 $^{\text{n.s.}}$
W: Morality of AI use	.40	3.22 **	-.03	-.26 $^{\text{n.s.}}$
M: Message credibility	-	-	.70	12.48 ***
X1*W	.22	1.35 $^{\text{n.s.}}$	.06	.44 $^{\text{n.s.}}$
X2*W	-.12	-.75 $^{\text{n.s.}}$	.12	.90 $^{\text{n.s.}}$
X3*W	.42	2.38 *	-.01	-.07 $^{\text{n.s.}}$
COV: Age	.01	.80 $^{\text{n.s.}}$	-.00	-.50 $^{\text{n.s.}}$
COV: Gender	-.13	-.91 $^{\text{n.s.}}$	.00	.02 $^{\text{n.s.}}$
COV: Industry type	.12	.79 $^{\text{n.s.}}$	.20	1.70 $^{⸸}$

Conditional indirect effect(s) of X (author types) on Y (attitude towards the company) at values of the moderator (M-1SD, M, M+1SD)

Bootstrap 95 percent confidence intervals for conditional indirect effects

	Morality	*b*	Lower	Upper
X1: Human supported by AI vs. Human	3.67	-.64	-1.19	-.09
	5.02	-.43	-.73	-.15
	6.37	-.23	-.53	.06

	Morality	*b*	Lower	Upper
X2: AI controlled by Human vs. Human	3.67	-.03	-.59	.43
	5.02	-.15	-.42	.11
	6.37	-.26	-.54	.03

X3: AI vs. Human	Morality	*b*	Lower	Upper
	3.67	-.96	-1.47	-.47
	5.02	-.56	-.86	-.27
	6.37	-.16	-.47	.17

Note. ⸸ $p < .1$, * $p < .05$, ** $p < .01$, *** $p < .001$, M = mean, SD = Standard deviation, n.s. = not significant.

Paper 5

Seeking empathy or suggesting a solution? Effects of chatbot messages on service failure recovery to optimize digital platform efficiency

Status: under review after revise-and-resubmit at the journal *Electronic Markets*

Co-Authors: Anna Rozumowski, Jan Freidank and Alexander Haas

Own Share: 70%

Earlier versions of this research project were presented as:

Haupt, M. and Rozumowksi, A. (2021), "Sorry I am Still Learning – Active Expectation Management of Chatbots, E-Society, 19th International Conference (virtual), March 3rd, 2021

Haupt, M. and Rozumowksi, A. (2021), Active Expectation Management in Chatbot Conversations, Australia and New Zealand Marketing Conference (ANZMAC), Melbourne, 2021, Dezember 1st, 2021.

Seeking empathy or suggesting a solution? Effects of chatbot messages on service failure recovery to optimize digital platform efficiency

Abstract

Chatbots as prominent form of conversational agents are increasingly implemented as a user interface for digital customer-firm interactions on digital platforms and electronic markets, but they often fail to deliver suitable responses to user requests. In turn, individuals are left dissatisfied and turn away from chatbots, which harms successful chatbot implementation and ultimately firm's service performance. Based on the Stereotype Content Model, this paper explores the impact of two universally usable failure recovery messages as a strategy to preserve users' post-recovery satisfaction and chatbot re-use intentions. Results of three experiments show that chatbot recovery messages have a positive effect on recovery responses, mediated by different elicited social cognitions. In particular, a solution-oriented message elicits stronger competence evaluations, whereas an empathy-seeking message leads to stronger warmth evaluations. The preference for one of these message types over the other depends on failure attribution and failure frequency. This study provides meaningful insights for chatbot technology developers and marketers seeking to understand and improve customer experience with digital conversational agents in a cost-effective way.

Keywords: Artificial intelligence, Chatbot, Service failure, Failure recovery, Social cognitions, Digital platform

1. Introduction

Driven by innovative technological advancements such as Artificial Intelligence or machine learning, chatbots are widely used nowadays and provide customer service on digital platforms such as social media, enterprise messengers or websites (Pizzi et al., 2021; Stoeckli et al., 2020). These agents increasingly substitute for human staff in electronic markets (van Pinxteren et al., 2020) and the global chatbot market is predicted to rise substantially from $17 billion in 2020 to over $102 billion in 2026 (Mordor Intelligence, 2021). As a remarkable and most recent example, Open AI's "ChatGPT" has attracted over 1 million users in five days, and is sought to disrupt numerous tasks in marketing, law or journalism and might even threaten Google by offering more humanlike answers and a smoother experience (Olson, 2022).

However, despite these technological advancements and considerable market potential, chatbots often fail in practice to deliver satisfactory responses to users' requests (Adam et al., 2020; Brandtzaeg & Følstad, 2018; Seeger & Heinzl, 2021). Customers are often left dissatisfied after receiving a response failure message from chatbots, which leads firms to risk negative consequences such as usage discontinuance and a decrease in firm performance (Diederich et al., 2020; Weiler et al., 2022). According to a recent survey from the banking industry, four out of five consumers is dissatisfied with chatbot interactions and almost 75% of consumers confirm that chatbots are often unable to provide correct answers (Sporrer, 2021). Concerning the consequences, about one third of consumers (30%) stated that they would turn away from the company or spread negative word of mouth after just one negative experience with the chatbot. Due to that threat and high levels of service failures, numerous companies including Facebook or SAP shut down chatbots on their digital platforms (Dilmegani, 2022; Thorbecke, 2022).

A chatbot response failure refers to an inadequate answer or no answer at all, which is sometimes also labelled as conversational breakdown (Benner et al., 2021; Weiler et al., 2022). Chatbot response failures reflect a service failure for the company, as the digital agent was unable to deliver satisfying information to support users’ goals. Users evaluate chatbots response failures as an insufficient service offer, comparable to response failures from a human frontline service employee, service robots, or other digital self-service technologies (Sungwoo Choi et al., 2021; Collier et al., 2017). Service failures have serious impacts on firms as they harm favorable customer reactions such as satisfaction, loyalty or positive word-of-mouth (Roschk & Gelbrich, 2014). According to research from Qualtrix and ServiceNow (2021), almost half of the respondents consider switching brands already after a single negative customer service interaction, and U.S. companies risk losing around $1.9 trillion of customer spending annually due to such poor experiences. This is particularly relevant for digital platforms (e.g., Airbnb, Uber), as their major basis for value creation resides in providing “efficient and convenient facilitation of transactions” (Hein et al., 2020, p. 91). In contrast to other industries, these platforms highly depend on their service offer (instead of products) and positive user experiences. Thus, providing the option to use chatbots offers large efficiencies for them, but at the same time also pose a threat in case of insufficient implementation.

Failure recovery strategies are therefore urgently needed to mitigate negative user responses and financial losses. In this regard, recovery messages are suggested as a viable option for chatbots to mitigate negative responses after self-inflicted response failure by the chatbots (Ashktorab et al., 2019; Benner et al., 2021). Such recovery messages aim to increase the chatbot's response capabilities to address the response failure, but also to mitigate negative user reactions to reduce the impacts of the perceived service failure.

Yet, relatively little is known about the impact of recovery strategies in chatbot conversations on customer responses. Recent studies have rather focused on reasons for response failures (Janssen et al., 2021; Reinkemeier & Gnewuch, 2022), identified different recovery strategies (Benner et al., 2021) or assessed user preferences for diverse recovery strategies (Ashktorab et al., 2019). These scholars asserted the potential of such strategies to prevent negative reactions following a failure (Benner et al., 2021; Reinkemeier & Gnewuch, 2022), but analyses of the effectiveness of recovery messages or comparisons of different types remain scarce. In contrast to such post-failure messages, Weiler et al. (2022) investigated how ex-ante messages (i.e., at the beginning of the chatbot interaction) influence users discontinuance of the chatbot interaction.

Based on the Stereotype Content Model (Cuddy et al., 2008) and the results of a pilot study which assessed chatbot's failure recovery strategies in real life, this study investigates the effects of two fundamental recovery message types – namely, seeking user's empathy versus suggesting a solution– on user's perceived warmth and competence, as well as on post-recovery responses. Furthermore, this research aims to understand under which circumstances which message type is advantageous regarding user satisfaction. Therefore, it considers two situational factors, in particular failure attribution and failure frequency. Results of three experimental studies show that both messages (i.e., empathy and solution) trigger specific social cognitions, more precisely either higher warmth or competence perceptions. In turn, these perceptions were found to influence people's post-recovery satisfaction and re-use intentions – but they do so to different degrees depending on the context.

This research contributes to the growing literature in information systems (IS) related to chatbots as digital conversational agents and offers relevant implications for firms how their chatbots should respond to a response failure in different contexts. Thereby, we integrate the technological (and IS) perspective related to chatbots' limited functionality and response failures with the service-oriented (and consumer psychology) perspective of recovery attempts to the service failure occurred. Our findings highlight the possibility to use recovery messages as low-cost, easy to program and universally usable strategy. Furthermore, they reveal the need to design a chatbot conversation carefully, and that the choice of an effective recovery message depends on situational factors. Recommendations for chatbot software developers and chatbot-employing firms are provided.

2. Conceptual background

2.1. Chatbots as digital conversational agents

Chatbots are text-based digital conversational agents that use natural language processing to interact with users (Gnewuch et al., 2017; Wirtz et al., 2018). These features lead to higher interaction and intelligence levels compared to other IS technologies (Maedche et al., 2019). Chatbots are a cost-effective tool for companies to automate customer-firm interactions while maintaining value and personalized service for their clients. Due to the convenient, easy and fast service and their 24/7 availability, the integration of chatbots is growing exponentially in various industries such as service, hospitality, healthcare or education (van Pinxteren et al., 2020). With the rise of chatbots, research increased tremendously in the last years, and scholars mainly investigated chatbot interactions from three perspectives, namely digital agent's design elements, (Diederich et al., 2020; Gnewuch et al., 2018; Kull et al., 2021), consumer responses to the digital interaction (Mozafari et al., 2022) and consumer responses to chatbot failures (see Sands et al., 2022 for an overview). Among these research fields, finding appropriate solutions for recovery of chatbot failures is particularly relevant, as it determines consumers' continuance decisions and ultimately a chatbot's success (Adam et al., 2020; X. Lv et al., 2022; Song et al., 2022). This is because, despite continuous development and the promising advantages for both customers and companies in service encounters, chatbots often do not live up to customer expectations and fail to understand or process user enquiries (X. Lv et al., 2022; Weiler et al., 2022; Xu & Liu, 2022).

2.2. Chatbot response failures

Lately, scholars have started to analyze the impacts of chatbot response failures. For instance, Seeger and Heinzl (2021) showed that digital agent's failures harm customer trust and stimulate negative word-of-mouth. Chatbot response failures also increase people's frustration and anger (Gnewuch et al., 2017; Mozafari et al., 2022; van der Goot et al., 2021), and create skepticism and reluctance to follow the bot's instructions (Adam et al., 2020). As a consequence, users frequently quit the conversation (Akhtar et al., 2019) and might even reject future chatbot interactions (Benner et al., 2021; van der Goot et al., 2021).

Chatbots fail frequently, because the processing of natural language input was found to be a complex task for machines due to unpredictable entries (Brendel et al., 2020). Moreover, chatbots are often integrated on digital platforms in wrong use cases and not connected to relevant data sources (Janssen et al., 2021; Mostafa & Kasamani, 2021). In addition, users were found to have exaggerated expectations of chatbots due to their human-like design. According to the 'computers are social actors' (CASA) paradigm, people ascribe social rules, norms and expectations to interactions with computers although they are aware that they are interacting with a machine (Nass et al., 1996). As such, people expect a chatbot to understand their request and respond with a suitable answer, just as they would expect of a human (Wirtz et al., 2018).

Parallel to the increased interest in chatbot technology, research on chatbot failure recovery strategies has gained traction in recent years (see Table 1 for an overview). This literature stream can be divided into three major sub-divisions. First, some scholars reviewed the literature or conducted expert interviews to derive critical success factors for chatbot interactions (Janssen et al., 2021) or categories of recovery strategies (Benner et al., 2021; Poser et al., 2021). The second body of research empirically examines how chatbot interaction could be designed pre-failure in order to mitigate negative consumer perceptions due to failures. For instance, research results indicate that higher chatbot anthropomorphism (Seeger & Heinzl, 2021; Sheehan et al., 2020), or specific message techniques (Weiler et al., 2022) positively influence consumer responses before the failure occurs. Third, and contrasting this, other scholars investigated the effects of post-failure recovery strategies. As one of the first studies, Ashktorab et al. (2019) compared user preferences of eight different recovery strategies and found that providing explanations or options of answers are favored as they display chatbot initiative. Mozafari et al. (2022) assessed that the mere disclosure of the chatbot (vs. human) identity has already a mitigating effect following failure. Further scholars found that chatbots are preferred over human agents after a functional failure (but not after a non-functional failure) (Xing et al., 2022), and chatbot self-recovery (vs. human agent recovery) leads to more positive user reactions (Song et al., 2022).

Scholars have also started to investigate effects of post-failure messages, and discovered for instance that some communication patterns (e.g., chatbot as ‘victim’ or ‘helper’) lead to more positive responses than other patterns (e.g., ‘persecutor’). (Brendel et al., 2020). Other studies revealed that cute or empathic responses (X. Lv et al., 2021; X. Lv et al., 2022), expressions of gratitude or apology (L. Lv et al., 2022) or self-depreciating humor (Xu & Liu, 2022; Yang et al., 2023) lead to more positive consumer reactions. Moreover, messages highlighting the human-chatbot relationship (i.e., appreciation message) were found to be more effective to increase post-recovery satisfaction compared to apology-related message strategies (Song et al., 2023).

2.3. Chatbots and the Stereotype Content Model

Following related studies about human-machine interactions (i.e., robots or chatbots), people quickly draw inferences about a bot’s personality as interaction partner similarly as they would evaluate a human frontline employee (Belanche et al., 2021; Sungwoo Choi et al., 2021). For example, following the Computers-are-Social-actors (CASA) paradigm (Nass et al., 1996), consumers are expected to evaluate a chatbot as digital interaction partner similarly as they would evaluate a human conversation partner – for instance by assessing its warmth and competence.

According to the Stereotype Content Model (Fiske et al., 2007) as one of the most established frameworks regarding social cognitions, people use *warmth* and *competence* as two universal dimensions of social perception when judging others. Thereby, warmth covers aspects like honesty, kindness or trustworthiness, while competence perceptions reflect capability, confidence, intelligence and skillfulness (Dubois et al., 2016; Fiske et al., 2007; Judd et al., 2005). Taken together, these dimensions are suggested to “account almost entirely how people characterize others” (Fiske et al., 2007, p. 77). Originally, this system of social judgment was applied to explain perceptions of social groups (Fiske et al., 2007) or individuals (Judd et al., 2005). Since then, scholars have extended its use to brands (Aaker et al., 2010) and more recently to service interactions with humans (Scott et al., 2013) or non-human entities (i.e., robots or virtual agents) (Sungwoo Choi et al., 2021; Kull et al., 2021; Xu & Liu, 2022).

Judgments of warmth and competence influence how people interact with others, as well as how people feel and behave (Cuddy et al., 2008; Marinova et al., 2018). Warmth is generally linked to cooperative intentions and prosocial behavior, whereas competence is associated with the power and ability to realize one’s goals (Cuddy et al., 2008). Inferred warmth and competence assessments enhance customer- and service-related outcomes such as satisfaction, trust or brand admiration, and they influence downstream behaviors like purchase intentions and retention (Aaker et al., 2010; Cuddy et al., 2008; Marinova et al., 2018; Scott et al., 2013).

Recently, scholars have increasingly investigated the impact of social cognitions, i.e., warmth and competence perceptions, on various outcomes in the field of digital agents (Belanche et al., 2021; Sungwoo Choi et al., 2021; Kull et al., 2021; McKee et al., 2022; Xu & Liu, 2022). These studies mainly focus on anthropomorphism effects. For instance, Sungwoo Choi et al. (2021) found that people perceive humanoid (vs. nonhumanoid) service robots as warmer but not as more competent. In turn, higher warmth influences satisfaction after a failure and supports recovery effectiveness. In contrast, Belanche et al. (2021) revealed that both dimensions of warmth and competence indicate a robot’s level of “humanness,” and both dimensions positively influence customers’ loyalty. Warmth and competence perceptions are also found to influence human-digital agent collaboration. More precisely, perceptions of these social cognitions predict people's choice of a particular agent, irrespective of the agent's objective performance level (McKee et al., 2022). Moreover, Xu and Liu (2022) found that humorous chatbot answers increase consumer's tolerance after a service failure, mediated by higher warmth and competence. Finally, Kull et al. (2021) found that when chatbots use a warm (vs. competent) initial message, people's brand engagement increased, because they feel closer to the brand in that condition. Despite these initial insights, however, little is known about effects of message-related cues on respondents' warmth or competence evaluations and subsequent service assessments. This gap is relevant because many chatbots are text-based agents, and thus, users mainly have to rely on the chatbot's (text-based) messages as cues to e.g., evaluate the chatbot's warmth and competence (van Pinxteren et al., 2020). Moreover, although chatbot service failures are common (Seeger & Heinzl, 2021), scholars confirm that there is still a lack of scientific knowledge about chatbot service recovery and its effectiveness (Xu & Liu, 2022). Therefore, this study evaluates how two distinct chatbot messages increase perceptions of social cognitions and enhance subsequent recovery responses.

2.4. Recovery strategies for chatbot failure

As chatbot response failures seem inevitable and lead to severe negative outcomes, firms are well advised to consider failure recovery strategies (Benner et al., 2021; Janssen et al., 2021). Thereby, a recovery strategy refers to an "effort [that] mitigates the previous negative effect of the failure" (Roschk & Gelbrich, 2014, p. 196). Scholars have revealed a wide range of such strategies as organizational responses, mainly with regards to service failures (for an overview, see van Vaerenbergh et al. (2019)). There are two basic dimensions of such failure recovery responses, namely (1) tangible compensation, such as monetary refunds and (2) psychological compensations, including positive service employee behavior (Roschk & Gelbrich, 2014; van Vaerenbergh et al., 2019). (1) Tangible compensations mainly include financial and process-related efforts within a firm. A common approach to tangible compensation in chatbot failures is to hand-over the conversation to a human employee to manage the problem and to prevent negative experiences (Ashktorab et al., 2019; Janssen et al., 2021). However, this solution comes with additional costs and reduces the level of automation (Reinkemeier & Gnewuch, 2022). In contrast, (2) psychological compensations generally come without costs and could be executed by the service encounter agent (i.e., frontline employee or chatbot) directly. Prominent examples are apologies from the service employee or expressions of regret for the occurred failure (van Vaerenbergh et al., 2019). This research focuses on psychological compensations, as this is of interest for both research and management: Scientifically, this study complements initial research which evaluates effects of different message elements (such as expressions of humor, cuteness, apology or gratitude) (L. Lv et al., 2022; X. Lv et al., 2021; Xu & Liu, 2022; Yang et al., 2023). Managerially, this type of compensation requires fewer resources (vs. human recovery) and can be integrated directly into the conversational process. In fact, a textual addition is all that is required to deliver these types of psychological compensation.

As gestures and nonverbal behaviors do not exist in chatbot conversations, people judge the chatbot conversation based mainly on written messages (van Pinxteren et al., 2020). We, therefore, analyze how different messages trigger social cognitions. As the study's outcome, post-recovery satisfaction and re-use intentions were chosen to evaluate recovery effectiveness. Post-recovery satisfaction represents one of the most widely used metrics to indicate successful recovery efforts (Song et al., 2022; Worsfold et al., 2007; Yang et al., 2023). Re-use intentions indicate continued acceptance of a chatbot and are relevant for its long-term success on digital platforms (Adam et al., 2020; X. Lv et al., 2022; Weiler et al., 2022).

Different messages as failure recovery strategies

Even small changes in the framing of communication messages were found to influence people's judgments and behaviors (You et al., 2020). Regarding chatbot conversations, different messages could be used in response to a service failure. In this research, two distinct message types labeled as an empathy-seeking message or solution-oriented message were deliberately chosen as they (1) represent the most common failure recovery strategies as revealed by our pilot study (see below in the empirical studies section) and (2) are thought to influence warmth and competence perceptions, respectively. Both types express a request from the chatbot. As first type, a chatbot might ask for a user's empathy and understanding regarding its limited abilities. This request for understanding is sought to elicit empathic concern for the chatbot's "infancy" and difficulties in handling requests. Scholars also refer to this message as "social" recovery strategy, which reflects apologizing for the failure "to appeal to the users' empathy and understanding similar to that which is shown in human-human conversations" (Benner et al., 2021, p. 9).

Empathy is defined as a person's intellectual or imaginative understanding of another person's condition or state (Hogan, 1969). Related to service, empathic customers were found to be less angry and more forgiving when they encounter a service failure (Wieseke et al., 2012). In a study with "classic" human frontline employees, customer empathy towards an employee was found to enhance social interactions, foster supportive attitudes and create a more satisfying experience (Lazarus, 1991; Wieseke et al., 2012). Scholars in the field of social service research support this, showing that empathy-related expressions are often beneficial to build or strengthen social bonds between interaction partners (Gerdes, 2011), which in turn increase warmth perceptions (Cuddy et al., 2008; Judd et al., 2005).

These well-documented effects could be also observed in human interactions with digital agents. As a chatbot reflects a digital version of a service employee, a chatbot message that evokes empathy (e.g., asking for patience and to hold on to the joint interaction) should trigger these warmth perceptions. Scholars consistently demonstrated that humans can feel empathy with inanimate objects such as chatbots or robots (Misselhorn, 2009). Related to the adjacent field of service robots, Wirtz et al. (2018) concludes that a bot's social-emotional and relational elements (e.g., social interactivity) increase warmth. As the chatbot's empathy message contains mainly such social-emotional and relational elements (e.g., asking for patience and to hold on to the joint interaction), we propose:

H1. The message type empathy increases consumer-perceived chatbot warmth.

As an alternative option, a chatbot could request the user to adapt the input to the chatbot's abilities, e.g., by rephrasing the input in short and simple words. Input from users was often found to be complex, and a shorter and more precise input has a higher probability of being processed correctly (Ashktorab et al., 2019; Luger & Sellen, 2016). Indeed, conversational agents were found to respond more successfully when the input was rather simple, short and unambiguous (Luger & Sellen, 2016). This type of request could be labeled as a solution-oriented message, as the chatbot tries to solve the failure actively. Related IS research has already used the solution-oriented message (i.e., "please rephrase your inquiry and try again") to encourage users for continuing with the chatbot (Benner et al., 2021; Weiler et al., 2022). While Weiler et al. (2022) use this message as ex-ante strategy at the beginning of the interaction, this study employs it as ex-post strategy to address the chatbot response failure directly when it occurred.

This concept has also been observed in human service interactions. When a frontline-employee focuses on the task (vs. social components) as the "core" of the service delivery and offers a possible solution to make the interaction more successful and convenient, this task-related behavior increases the perceived competence of this employee (Marinova et al., 2018). Several scholars support this argumentation, and acknowledge that competence-oriented messages imply that service providers are "very capable in providing consumers with solutions" (Huang & Ha, 2020, p. 620).

Related to the chatbot, the message-type solution focuses on the task, that is, to make the interaction with the customer effective. As consumers perceive digital assistants such as chatbots as social actors (Nass et al., 1996; van Pinxteren et al., 2020), this solution-oriented message should increase chatbot competence perceptions (Marinova et al., 2018). . Moreover, the solution message indicates that the chatbot is aware of the linguistic complexity of user input and of options to improve the quality of the chatbot's answer (Weiler et al., 2022). Both aspects (i.e., awareness of a problem, and presentation of a possible solution) indicate a kind of skillfulness or intelligence, two key items reflecting competence (Cuddy et al., 2008; Xu & Liu, 2022). In addition, related service robot literature proposed that a when a bot can serve a user's functional needs (e.g., offering a solution to a request), this service enhances perceptions of its usefulness and competence (Wirtz et al., 2018). Therefore, we hypothesize:

H2. The message type solution increases consumer-perceived chatbot competence.

According to scholars, warmth and competence perceptions can serve as underlying mechanisms that explain how consumers respond to technology infusion in service (Belanche et al., 2021; van Doorn et al., 2016). According to van Doorn et al. (2016), warmth and competence perceptions elicited by digital service technology both enhance consumers' satisfaction and loyalty intentions. A chatbot-study found that if chatbots could elicit warmth perceptions within human-chatbot interactions, chatbot use is rising (Mozafari et al., 2021). Supporting this, research from the related field of service robots found that warmth perceptions significantly increased post-failure satisfaction and loyalty (Sungwoo Choi et al., 2021). Similarly, research confirmed that consumers' competence perceptions (e.g., the belief that chatbots are capable to fulfill a task or enable successful service recovery) increase their interaction satisfaction and re-use intentions (X. Lv et al., 2022; Mozafari et al., 2022). Further studies about human (Babbar & Koufteros, 2008; Güntürkün et al., 2020; Habel et al., 2017) and digital service agents (Belanche et al., 2021) support that higher warmth and competence perceptions drive consumers' service value perceptions, satisfaction and loyalty. Thus:

H3. Stronger consumer-perceived (a) warmth and (b) competence increase consumers' post-recovery satisfaction.

H4. Stronger consumer-perceived (a) warmth and (b) competence increase consumers' chatbot re-use intentions.

2.5. Factors influencing the perception of recovery messages

Research has shown that situational factors regarding chatbot interactions influence user perceptions and responses (Gnewuch et al., 2017; Janssen et al., 2021; Pizzi et al., 2021). Therefore, we identified two relevant factors, namely *failure attribution* and *failure frequency*, which are thought to impact users' reactions and preference for one of the recovery messages. Both factors were found to be important elements in the failure and recovery literature (Sunmee Choi & Mattila, 2008; Collier et al., 2017; Ozgen & Duman Kurt, 2012; van Vaerenbergh et al., 2019).

Failure frequency

In chatbot conversations, users regularly need to make multiple attempts to enter a request in a way that the chatbot will understand (Ashktorab et al., 2019). That means that many initial service failures are not recovered adequately but lead to a second service failure – a situation also labeled as *double deviation* (Johnston & Fern, 1999; van Vaerenbergh et al., 2019). Such double deviations were found to reinforce negative customer responses that were caused by the first failure, such as customer dissatisfaction, anger, or churn (Ozgen & Duman Kurt, 2012; van Vaerenbergh et al., 2019). Furthermore, people were found to prefer different recovery strategies for a single vs. double deviation, leading to the conclusion that the service provider should adequately account for the failure frequency in choosing the appropriate recovery strategy (Pacheco et al., 2019). Therefore, chatbot creators need to identify the best-possible "match" for the response to the failure (Roschk & Gelbrich, 2014). After a first failure, both response messages are expected to mitigate negative consequences via the paths of warmth and competence as proposed above. Yet, when users re-enter their request and the chatbot fails again to deliver an appropriate answer, this represents a new situation with (potential) implications for the effectiveness of both message types after the second failure.

The empathy-related message seeks to evoke understanding and empathy and create feelings of warmth and mutual connection (Cuddy et al., 2008; Lazarus, 1991; Wieseke et al., 2012). Asking for understanding regarding the chatbot's limited abilities is possible at any interaction stage or situation, as the chatbot refers to its own lack of abilities (vs. the user). Therefore, an empathy message is assumed to create warmth perceptions irrespective of the failure frequency. Related to the message type solution, as argued above, people are expected to accept the request to re-phrase their input to better adhere to a chatbot's needs after a first failure and even perceive that chatbot as competent (Chong et al., 2021; Marinova et al., 2018). However, after re-phrasing the request and being confronted with a second service failure, this competence perception is assumed to be negatively affected as the chatbot was not able to provide a solution. As Johnston and Fern's (1999) study showed, more than half of the respondents lost confidence in a service agent's competence after a double deviation. Taken together, after a double deviation, empathy-seeking message should be more effective than solution-oriented messages. Formally:

H5. After a double deviation, an empathy message is more effective than a solution message in that the effect of the empathy message on consumer-perceived chatbot warmth is stronger than the effect of the solution message on consumer-perceived chatbot competence.

Failure attribution

Following attribution theory (Weiner, 1985, 2012), particularly in its application to service failures, customers seek to attribute the responsibility for the occurrence of a negative incident to some person or thing as a way to understand the situation and regain control over their environment. Thereby, people mainly differentiate between two dimensions of a so-called "locus of control" – either they blame others (i.e., external attribution) or they blame themselves (i.e., internal attribution) for the failure that has occurred (Weiner, 1985). Previous research showed that customers respond differently to service failures depending on which party they believe to be responsible for the failure (Sunmee Choi & Mattila, 2008; Collier et al., 2017). For instance, when people assign the firm or its service agent as responsible for the failure, people react more negatively than when they perceive that they are (at least partially) responsible for the failure as well (Sunmee Choi & Mattila, 2008). Consequently, people respond more positively to service failures that are self-attributed (versus firm-attributed) remain more satisfied with the firm and are more likely to forgive such failures (Sunmee Choi & Mattila, 2008; Gelbrich, 2010).

When considering which chatbot recovery message should be employed (i.e., solution or empathy), failure attributions are supposed to differentiate its effectiveness. More precisely, we expect that the failure attribution and the recovery message should match the failure type to create positive outcomes. Recovery research has shown that matching the recovery strategy with the failure type (e.g., monetary compensation for monetary failure) is more effective than a non-match (Roschk & Gelbrich, 2014). Related to chatbot interaction, when users attribute the failure to the chatbot (i.e., blame it for the failure), an empathy (vs. solution) message should be a better match, as in that case the attributed party "takes the blame" by asking for empathy and understanding. Scholars have established that such messages send cues that clarify and acknowledge blame attributions, and they help users to understand the possible reason for the failure (e.g., the "infancy" of the chatbot). In turn, these cues work as a coping mechanism to handle the negative consumer reactions caused by the failed service (Gelbrich, 2010). In line with that, an empathy message as response to a chatbot-caused failure is supposed to match, while the solution message expresses that the user is also part of the failure – a message cue which does not match the responsibility perception of the user.

Vice versa, the solution message matches a user-attributed failure because it offers guidance for the user to tailor the request to the chatbot. When a user acknowledges to be (at least partly) responsible for the failure or is unsure about who to blame, a solution message (vs. empathy) should better match this perception. To put it differently, users are supposed to accept a request to rephrase their entry when they admit to be part of the problem (Sunmee Choi & Mattila, 2008), and they might even be thankful for guidance on how to react in the interrupted process. Yet, when a chatbot is believed to be the responsible party, a solution message that expresses a user action to resolve the situation is expected to be perceived as less appropriate, and should therefore affect consumers' competence perceptions to a smaller extent. Thus:

H6a. An empathy message leads to higher consumer-perceived warmth in the case of a chatbot-attributed failure (match) than in the case of a user-attributed failure (mismatch).

H6b. A solution message leads to higher consumer-perceived competence in the case of a user-attributed failure (match) than in the case of a chatbot-attributed failure (mismatch).

Figure 1 illustrates our conceptual model.

[Insert Figure 1 about here]

3. Empirical studies

3.1. Pilot study

As initial pilot study, chatbots from different companies and across industries in the DACH-region (Germany, Austria, and Switzerland) were analyzed to assess which recovery strategy they used after a service failure. A service failure reflects that a chatbot did not understand the user's request and was provoked by entering some random letters as incomprehensible input. The final sample resulted in 101 chatbots from business, education and public administration. Almost a third of these bots (i.e., 27) did not allow any free-text entry but only a set of options to choose, and consequently no "failure" in communication could occur when engaging with them. Out of the remaining 74 (free-text processing) chatbots, 34 ask the user to reformulate their request, reflecting the solution message type. Users were asked to use short sentences, simple words, and to be as precise as possible in their wording. Furthermore, 12 chatbots appealed to the user's empathy and understanding. Lastly, no clear strategy was identified for 28 chatbots, and most of these chatbots just replied with a simple error feedback message. That means the chatbot just sends short messages like "Sorry I did not understand that."

In sum, the pilot study revealed that four major message-based recovery strategies are prominent in chatbot conversations: (1) pre-defined answers, (2) a solution-oriented message, (3) an empathy-seeking message, and (4) a simple error feedback message. As pre-defined answers limit the variety of entries, they are generally less flexible. Therefore, this message type was omitted and the latter three types were analyzed.

3.2. Study 1

Study design

To investigate the influence of type of the recovery messages on users' post-recovery satisfaction, study 1 applies a one-factorial between-subjects experiment with three cases (message type: empathy vs. solution vs. control). Participants were recruited from two

European universities through email distribution lists and randomly assigned to one of the scenarios (see Figure 2 for detailed scenarios). After excluding four participants who failed in the attention check (i.e., “If you read this, please press button 1”), our sample resulted in 178 participants (M_{Age}: 24 years, SD_{Age}: 18.34, 56.2% females). Participants had to imagine that they interact with a chatbot of an electronics provider, as electronic retail and service offers are nowadays vastly provided by digital platforms and electronic markets, and prior research considered e-commerce as prevalent field of chatbot service (Adam et al., 2020; Alt, 2020; Gnewuch et al., 2017).

[Insert Figure 2 about here]

As for the conversation, three questions about a camera were asked; two of which the chatbot answers correctly and the last one where the chatbot mentioned a non-understanding of the user request (i.e., response failure, see Table 2). As a manipulation, we varied the failure responses: The chatbot either asked the user to have empathy with its limited abilities and to try again (i.e., type empathy), or to adapt and simplify the input (i.e., type solution). As control case, the chatbot just replied, “Sorry, I did not understand your request.” As manipulation checks, we relied to Hosseini and Caragea (2021) as they described empathy-seeking behavior: People in the empathy message scenario perceived more strongly that the chatbot had “asked for their empathy and understanding” ($M_{Empathy}$: 6.31, $M_{Solution}$: 2.88, $M_{Control}$: 3.02; $F = 90.14, p < 0.001$) compared to the other scenarios. Likewise, for the solution case we relied on Marinova et al. (2018) to describe problem-solving behavior: Respondents of the solution message perceived more strongly that the chatbot “has asked to rephrase my request” ($M_{Empathy}$: 2.53, $M_{Solution}$: 5.98, $M_{Control}$: 2.32, $F(2,175) = 94.94$, $p < 0.001$). Thus, the manipulation was effective. Moreover, the scenarios were perceived as realistic (i.e., “The scenario is realistic” and “I can imagine a chatbot interaction happening like this in real life.” (α: 0.89, M: 5.71, SD: 1.41 (on a 7-point Likert scale)).

Measures

For all three studies, reflective multi-item measures with 7-point Likert scales (1= strongly disagree and 7 =strongly agree) from the extant literature were used and adapted to the study context. Post-recovery satisfaction was captured with three items from Agustin and Singh (2005). Perceived competence and warmth of the chatbot were captured by three item-scales each from Aaker et al. (2010), followed by some demographics (i.e., age, gender). Reliability and validity values were all above the thresholds (see Table 3). Cronbach’s alpha and composite reliability values are above the cut-off value of 0.70, indicating construct-level reliability (Hulland et al., 2018). Second, the average variance extracted (AVE) for every multiple-item construct exceeded 0.50, showing appropriate convergent validity. Third, the AVE values were found to be larger than the shared variance of any other remaining construct, indicating discriminant validity (Hulland et al., 2018). All items and factor loadings are illustrated in Table 3, and means and standard deviations for the main variables are provided in Table 4.

Results

An ANOVA revealed significant effects of the three message types on post-recovery satisfaction (F(2,175) = 15.97, $p < 0.001$). Post-hoc tests (Bonferroni) showed that both the empathy message and the solution message led to significantly higher post-recovery satisfaction than the control message ($M_{Solution}$: 2.93 vs. $M_{Control}$: 1.79, $p < 0.001$; $M_{Empathy}$: 2.60 vs. $M_{Control}$: 1.79). In contrast, the empathy and solution messages did not lead to significantly different satisfaction (p = 0.36). Thus, both messages enhance satisfaction compared to control – but not to a different degrees when compared to each other.

To test H1 to H3, a mediation analysis was conducted with PROCESS Model 4 using 5,000 bootstrapping samples and 95% confidence intervals (CIs) (Hayes, 2018). The message types were used as a multicategorical independent variable, warmth and competence served as parallel mediators, satisfaction was the outcome, and age and gender were covariates.

As hypothesized, the empathy message (vs. control) increased warmth ($b = 1.01$, $p < 0.001$), and the solution message (vs. control) led to higher competence perceptions ($b = 0.88$, $p < 0.001$), supporting H1 and H2 respectively. The empathy message (vs. control) did not increase competence perceptions (p = 0.94), while the solution message also increased warmth ($b = 0.54$, $p < 0.05$). In turn, both warmth ($b = 0.12$, $p < 0.05$) and competence ($b = 0.42$, $p < 0.001$) had a positive effect on satisfaction, supporting H3 (a and b). The indirect effects of the empathy message on satisfaction were significant via warmth ($b = 0.12$, [0.01, 0.27]), and they were significant for the solution message on satisfaction via competence ($b = 0.37$, [0.17, 0.60]).[1]

3.3. Study 2 – Failure frequency

Design and procedure

Study 2 examined the effect of the recovery messages on post-recovery satisfaction and re-use intentions under different failure recovery conditions (i.e., success vs. second failure after the

---

[1] We also conducted a study (students from two European universities, n=270, M_{Age}= 27 years, 52% female) with the same measures based on a further scenario (i.e., a chatbot as pizza delivery agent as food delivery represents another common field for digital platforms (e.g., Uber eats, Deliveroo, HelloFresh) and for chatbot services Li et al. (2020); van Pinxteren et al. (2020). Results of a mediation analysis (PROCESS model 4) showed that the empathy message (vs. control) led again to higher warmth perceptions ($b = 1.16$, $p < 0.001$) while the solution message (vs. control) did not ($p = 0.45$). The solution message (vs. control) led to higher competence perceptions ($b = 0.56$, $p < 0.05$), whereas the empathy message (vs. control) did not ($p = 0.56$). In turn, satisfaction was influenced by warmth ($b = 0.30$, $p < 0.001$) and competence ($b = 0.48$, $p < 0.001$). Neither message influenced satisfaction directly. In sum, the results also provide support for H1-H3 again and add further validity to Study 1.

recovery). Respondents were recruited via the online platform Prolific (U.S.-participants with 95% former tasks approval ratio). Participants were randomly assigned to a 3 (message type: empathy vs. solution vs. control) x 2 (recovery outcome: success vs. second failure) between-subjects experiment. Respondents had to imagine a chatbot interaction for a table reservation in a restaurant (see Table 2). The chatbot did not understand the initial user request and responded with one of the three message types from Study 1. After reading the recovery message, respondents had to rate their warmth, competence, anger and satisfaction and enter an individual input as response. On the next page, the survey tool illustrates the interaction including the individual user input and adds either a success message (i.e., "I successfully booked a table") or a second failure message. In case of the second failure, one of the three message types (i.e., empathy-seeking, solution-oriented, control) was displayed (again); with a slightly adapted text for the solution-message to fit the context. After these messages, respondents again rated their perceptions (i.e., warmth, competence, satisfaction, re-use intentions, anger). To increase realism and the fit of user-entry and message, we excluded fourteen participants in the recovery success condition who entered nonsensical input. Furthermore, we excluded seven participants who failed the attention check (i.e., participants who agreed to the false statement "the chatbot has forwarded me to a human service employee"), the final sample consisted of 237 respondents (M_{Age}: 45 years, SD_{Age}: 14.56, 49% female).

Measures

Scales were identical to those used in study 1 above. Chatbot re-use intentions were measured with the scale from Wallenburg and Lukassen (2011). As control variable, we assessed participant's anger with three items from Xie et al. (2015), as this emotional response could influence user reactions in chatbot interactions (Crolic et al., 2021). All scales displayed adequate validity and reliability (see Table 3). Moreover, scenarios were perceived as realistic

(M: 5.41, SD: 1.45) and the manipulation checks were effective. The empathy message was perceived as stronger for seeking empathy and understanding (M_{Solution}: 3.26, M_{Empathy}: 5.70, M_{Control}: 2.25, $F(2,234) = 71.16$, $p < 0.001$), and respondents in the solution message scenario agreed more that the chatbot has asked to rephrase the input as possible solution (M_{Solution}: 6.62, M_{Empathy}: 2.45, M_{Control}: 2.56, $F(2,234) = 155.62$, $p < 0.001$). Regarding the recovery success manipulation, participants in the success-scenarios (vs. second-failure) rated significantly stronger that the chatbot "has successfully reserved a table" (M_{Success}: 6.67, $M_{\text{Second-Failure}}$: 1.28, $t(235) = 50.82$, $p < 0.001$). Moreover, Table 4 provides descriptives for the main variables.

Results

To test H1 to H3 in one comprehensive model, we again conducted a mediation analysis (PROCESS Model 4, Hayes (2018) with 5,000 bootstrap samples and 95% CIs) with the same setup as in Study 1. Anger, age and gender were added as covariates. Consumer perceptions were evaluated after the first failure. The empathy message (vs. control) increased perceived warmth ($b = 1.22$, $p < 0.001$), and the solution message (vs. control) led to higher competence perceptions ($b = 0.49$, $p < 0.05$), supporting H1 and H2. The solution message also increased perceived warmth ($b = 0.50$, $p < 0.05$), whereas the empathy message did not increase competence ($p = 0.84$). Satisfaction was influenced by warmth ($b = 0.26$, $p < 0.001$) and competence ($b = 0.51$, $p < 0.001$), supporting H3. The indirect effect of the empathy message on satisfaction via warmth was significant ($b = 0.31$, [0.15, 0.51]) and the indirect effect of the solution message on satisfaction via competence was significant ($b = 0.25$, [0.02, 0.49]). In our analysis, neither of the two message types had a direct impact on satisfaction.

Next, to examine effects of responses to the second failure (H3, H4 and H5), we applied the initial mediation analysis and compared the different messages and evaluations following the second-failure condition.

Related to the effects of the mediators on the dependent variables (i.e. H3, H4), results of the mediation model with satisfaction showed that warmth and competence significantly increased post-recovery satisfaction ($b_{\text{warmth}} = 0.12$, $p < 0.05$; $b_{\text{competence}} = 0.79$, $p < 0.001$), supporting H3. Similarly, when using re-use intentions as dependent variable, warmth and competence significantly increased re-use intentions ($b_{\text{warmth}} = 0.28$, $p < 0.005$; $b_{\text{competence}} = 0.50$, $p < 0.001$), supporting H4.

Results of the messages after the second failure on the mediators (H5) show that the empathy message still led to perceived warmth ($b = 0.96$, $p < 0.005$), whereas the solution message did not lead to higher competence perceptions ($p = 0.13$). Thus, H5 could be supported. Correspondingly, the indirect effect of the empathy message on satisfaction via warmth was significant ($b = 0.11$, [0.01; 0.25]), whereas the indirect effect of the solution message on satisfaction via competence was not significant ($b = 0.35$, [-0.14; 0.87]). Similarly, the indirect effect of the empathy message on re-use intentions via warmth was significant ($b = 0.27$, [0.06; 0.56]), whereas the indirect effect of the solution message on re-use intentions via competence was not significant ($b = 0.22$, [-0.08; 0.63]). Thus, as hypothesized, the empathy message was found to be more effective than the solution message after the second failure.

We also evaluated respondents' evaluation of the three recovery messages in case of a successfully resolved second attempt. In this case, the empathy message had only a marginally significant impact on warmth ($b = 0.62$, $p < 0.1$), whereas the solution message had no impact on competence ($p = 0.35$). Warmth and competence increased post-recovery satisfaction ($b_{\text{warmth}} = 0.22$, $p < 0.001$; $b_{\text{competence}} = 0.56$, $p < 0.001$). Both indirect effects were not significant (i.e., empathy on satisfaction via warmth ($b = 0.14$, [-0.01; 0.39]); solution on satisfaction via competence ($b = 0.15$, [-0.14; 0.45]). Thus, message effects dissolve when the chatbot solved the user's request.

3.4. Study 3 – Failure Attributions

Design and procedure

Study 3 aimed to examine the effect of the recovery messages on post-recovery satisfaction and re-use intentions under different failure attribution conditions, i.e., either chatbot or user were responsible for the failure. Respondents from a German university were recruited via E-Mail distribution lists and randomly assigned to a 3 (message type: empathy vs. solution vs. control) x 2 (user fault vs. chatbot fault) between-subjects experiment. After excluding eight participants who failed the attention check (i.e., if you read this, please press button 1), the final sample consisted of 249 respondents (M_{Age}: 27 years, SD_{Age}: 14.24, 63% female). As scenario, a pizza delivery case was used (see Table 2), as this case represents another common field for digital platforms (e.g., Uber eats, Deliveroo, HelloFresh) and for chatbot services (Li et al., 2020; van Pinxteren et al., 2020).As user-fault scenario, the user entered “to my home” as the delivery address, which obviously could not be found in a database. As chatbot-fault scenario, the user entered an address “to Schlösschen Street 12,” which a chatbot would be supposed to find in a location database. Recovery messages were taken from Study 1 and slightly adapted to fit the failure situation.

Measures

After reading the scenario, participants rated their post-recovery satisfaction, followed by demographics and manipulation and realism checks. Scales were identical to the ones used in Study 1 and 2. All scenarios were perceived as realistic ($\alpha = 0.81$; *M*: 5.71, *SD*: 1.26). As manipulation check for failure attribution, respondents rated “who was responsible for the failure,” anchored at “user (1)” up to “chatbot (7).” People in the chatbot-fault scenario held the chatbot more responsible for the failure compared to the user-fault scenario ($M_{\text{Chatbot-fault}}$: 4.69; *SD*: 1.89 vs. $M_{\text{User-fault}}$: 3.56, *SD*: 2.27, $t(247)= -4.08$, $p < 0.001$). Moreover, for the message types, respondents of the empathy scenario rated significantly stronger that the chatbot asked for their empathy and understanding ($M_{Empathy}$: 5.33, $M_{Solution}$: 2.90, $M_{Control}$: 2.25, $F(2,246) = 130.93$, $p < 0.001$). Similarly, respondents in the solution message scenario perceived more strongly that the chatbot has suggested a solution ($M_{Solution}$: 4.34, $M_{Empathy}$: 3.10, $M_{Control}$: 2.39, $F(2,246) = 26.62$, $p < 0.001$). Again, all scales exhibited adequate validity and reliability (see Table 3). In addition, Table 4 shows the means and standard deviations of the key variables.

Results

To test H1 to H3 in one comprehensive model, we used a moderated mediation analysis (PROCESS Model 8, Hayes (2018) with 5,000 bootstrap samples and 95% CIs) with the same setup as in the studies above, including age and gender as covariates.

Regarding H1 and H2, results confirmed Study 1 and 2. Again, the empathy message (vs. control) increased perceived warmth *(b* = 1.18, *p* < 0.01), and the solution message (vs. control) led to higher competence perceptions *(b* = 1.22, *p* < 0.01), supporting H1 and H2. In addition, results showed that, the solution message (vs. control) did not increase warmth (*p* = 0.51) and the empathy message did not increase competence (*p* = 0.50). Satisfaction was influenced by warmth *(b* = 0.16, *p* < 0.01) and competence *(b* = 0.58, *p* < 0.001), supporting H3 again. Both message types had no direct impact on satisfaction.

Regarding H6a, the interaction of empathy message x failure attribution had no significant impact on warmth (*p* = 0.64). The indirect effect of the empathy message (vs. control) on satisfaction via warmth was significant in the case of a user-attributed failure *(b* = 0.19; [0.03, 0.41]) and in the case of a chatbot-attributed failure *(b* = 0.15; [0.04, 0.31]). Subsequently, the moderated mediation effect was not significant *(b* = -0.04; [-0.23, 0.14]). This means, irrespective of the failure attribution, there is a mediation effect of empathy on satisfaction via warmth. As a consequence, 6a could not be supported.

However, the situation changes when considering the solution message (H6b). In this case, the interaction of the solution message x failure attribution had a negative impact on competence *(b* = -1.23, $p < 0.05$). The indirect effect of the solution message (vs. control) on satisfaction via competence was significant in the case of a user-attributed failure *(b* = 0.71; [0.32, 1.14), but not significant in the case of a chatbot-attributed failure *(b* = 0.06; [-0.31, 0.45]). The index of moderated mediation was significant and negative *(b* = -0.65; [-1.22, -0.12]). This indicates that the positive (mediated) effect of the solution message through competence on satisfaction is only supported when the failure is attributed to the user. When the chatbot is responsible for the failure, the positive effect diminishes. In sum, H6b could be supported.

Finally, to test H4 (a and b), we applied the same moderated mediation model (Model 8) and replaced satisfaction with re-use intentions. Results are comparable to those above. Empathy led to warmth (*b* = 1.18, $p < 0.01$) and solution increased competence (*b* = 1.22, $p < 0.01$). Moreover, solution did not lead to warmth (p = 0.51) and empathy did not lead to competence (*p* = 0.50). 'In turn, chatbot re-use intentions were influenced by warmth (*b* = 0.16, $p < 0.05$) and competence (*b* = 0.69, $p < 0.001$), supporting H4 (a and b). The effects of moderated mediation remained comparable to those above: The indirect effects of empathy via warmth on re-use intentions were significant irrespective of failure attribution ($b_{\text{user-attribution}}$ = 0.19, [0.01; 0.46] and $b_{\text{chatbot-attribution}}$ = 0.15, [0.01; 0.33]; index = -0.04; [-0.25;0.16)) , while the indirect effects of solution via competence on re-use intentions were only significant in case of user-attribution (and not for chatbot-attribution) (i.e., ($b_{\text{user-attribution}}$ = 0.84, [0.37; 1.37] and $b_{\text{chatbot-attribution}}$ = 0.07, [-0.38; 0.54]; index = -0.77; [-1.48; -0.13]).

4. Discussion

As response failures occur frequently during chatbot interactions, recovery strategies are greatly needed to mitigate negative user reactions, avoid financial losses and assure re-use intentions.

This is especially relevant for electronic markets and digital platforms such as Airbnb, Booking or Uber, as service provision and customer-facing support are part of their key assets. To help answer the question of whether and how recovery messages might support these goals, the present research investigated how people respond to two characteristic recovery messages in chatbot conversations and focused on the mediating role of social cognition. Three experiments in two contexts compared the two characteristic messages empathy and solution and identified that these messages trigger social cognitions of warmth or competence (*H1* and *H2*) – which positively influence post-recovery satisfaction and chatbot re-use intentions *(H3 and H4)*. Furthermore, the impacts of situational factors on message effectiveness were analyzed. First, failure frequency also determines which message should be preferred (*H5*). More precisely, after a double deviation, only an empathy message has a positive effect on post-recovery satisfaction, while a solution message is less accepted and creates lower satisfaction. Second, integrating the factor of failure attribution (*H6a/b*) showed that a solution message is particularly detrimental to user satisfaction with a chatbot-attributed failure (i.e., a mismatch). In this situation, the solution message did not lead to higher satisfaction (mediated via competence). In contrast, in a user-attributed failure situation, people seemed to accept a solution message more, as this message type led to higher post-recovery satisfaction via increased competence perceptions. An empathy message was found to be acceptable for both user- and chatbot-failure attributions. This indicates that an apology and request for understanding is "always possible" and a less critical approach compared to the solution message, and rather preferable when failure attribution remains unclear. Finally, when the chatbot solved the user request successfully after an initial failure, effects of different recovery message dissolved. Thus, the success of an interaction shifts post-hoc perceptions of the previous recovery messages.

Theoretical contributions

This research responds to scholarly calls for further user-centered investigation of chatbot response failures (Diederich, Brendel, and Kolbe 2020) and provides several theoretical contributions. First, we add to the growing body of research regarding digital agents' conversational design (Crolic et al., 2021; Sands et al., 2021; Song et al., 2022; Weiler et al., 2022). Interactions in electronic markets and particularly digital platforms (e.g., Airbnb, eBay) rise continuously, leading to a parallel increase in demand for effective and efficient customer service (Hein et al., 2020; Suta et al., 2020). Next to such user-facing platforms, chatbots are also increasingly implemented in corporate applications (e.g., Slack or Microsoft Teams) to facilitate processes and information access (Stoeckli et al., 2020). Thus, as chatbots are increasingly taking over tasks in the digital surrounding and are a major service innovation, an appropriate design of chatbot responses is key for positive customer experiences and firm profitability (Mozafari et al., 2022). This study proposes that message types, when used as a psychological recovery attempt, should be carefully chosen depending on situational factors like failure frequencies or failure attribution. These results offer a more nuanced view on the effectiveness of recovery messages – and confirm former studies that stated that chatbot designs should follow human service chat interactions in order to be successful (Belanche et al., 2021; Gnewuch et al., 2018; van Pinxteren et al., 2020).

Second, this research adds to the literature of service failures and recovery, particularly in the domain of digital agents (Chong et al., 2021; Mozafari et al., 2022). With this study, we respond to scholars who have called for an examination of effective recovery strategies to improve users' service experience after chatbot failures (Benner et al., 2021; Janssen et al., 2021; van der Goot et al., 2021). We also complement the findings of Weiler et al. (2022) who examined ex-ante strategies by showing that messages directly after the failure (ex-post) also have a positive effect on re-use intention and thus reduced discontinuance. Moreover, this research complements studies which consider the impact of recovery messages of digital agents

(L. Lv et al., 2022; Song et al., 2023). As service delivery by chatbots becomes more widespread, understanding how people respond to chatbot recovery attempts is of crucial relevance to secure service quality and consumer loyalty (Mozafari et al., 2022; Sands et al., 2021). Supporting findings from related studies (such as Xu and Liu (2022), our study results show that messages could trigger different social cognitions and achieve their goal of increasing post-recovery satisfaction via different paths. In addition, this study examines several conditions that influence the effectiveness of a particular message. By including failure frequency (i.e., double deviation) and failure attributions in the research design, we illustrated that such dimensions indeed play a role for the optimal message choice. As such, this paper also adds to the scant research around double deviations (Pacheco et al., 2019) and to knowledge of the effects of failure attributions in the field of human-computer interaction. Additionally, results might encourage future related work to incorporate these factors into their research as well.

Third, this research adds to research assessing social cognitions. Only recently have scholars started to assess the perceptions of warmth or competence in relation to digital (conversational) agents (Sungwoo Choi et al., 2021; McKee et al., 2022; Xu & Liu, 2022). As new technology, such as Artificial Intelligence or machine learning, further develops, digital agents will interact in more human-like service interactions and will increasingly imitate human behavior in order to create more favorable user responses. While related studies in this field mainly concentrate on anthropomorphism as visual cues for warmth or competence (e.g., Sungwoo Choi et al., 2021), our research focused on text-related cues. This study uncovered warmth and competence perceptions as underlying effects of messages on user's post-recovery response. Message elements requesting a person's understanding are social-oriented and were perceived as warm, whereas a message which presents a possible solution is task-oriented and was perceived as competent. In turn, both perceptions increased post-recovery satisfaction. This supports the 'computers as social actors' (CASA) paradigm (Nass et al., 1996) and shows that chatbot responses are processed and perceived like human service-agent messages. However, the study also shows that the mediation through social perceptions could be eliminated by external circumstances. For instance, a double deviation (i.e., a chatbot's second non-understanding) removed the mediated effect of solution-oriented messages via competence.

Managerial implications

Results of the three studies provide guidance to both software designers and to companies employing chatbots on how to implement chatbot recovery messages as cost-effective and universally usable tool to mitigate negative service experiences. First, using a dedicated recovery message is beneficial to mitigate negative user's responses after a chatbot failure with only marginal costs for software programming. This research revealed that each message follows a distinct path to increase post-recovery satisfaction – either by driving competence-perceptions or warmth-perceptions of users. Uncovering these underlying mechanisms helps managers to understand how consumers' responses are formed. In particular, software designers can now formulate precise warmth- or competence-related messages as effective response to service failures.

Second, across the studies, competence perceptions generally exerted a stronger total effect on satisfaction than warmth. As the solution message fosters competence perceptions, this message type could therefore be considered as a more effective strategy for both product- and service-related contexts. Using the solution message also allows chatbot designers to employ corrective measures to successfully conclude the conversation. However, if the recovery process was successful after the initial failure (i.e., the chatbot successfully resolved the request), the impact of the recovery messages dissolved, as consumers do not seem to care (post-hoc) how they got to this point. Nevertheless, as likelihood of failure is high, managers and chatbot developers should be encouraged to incorporate one of the two message forms to safeguard against negative effects in case of failure without risking negative effects in case of success.

Third, the analysis of situational factors revealed that using the solution message is less effective after a double deviation. When failing twice, the empathy message led to warmth and satisfaction, while the solution message did not increase competence nor satisfaction. Thus, the solution message should be avoided in this situation. The empathy message is also preferable when people attribute the chatbot as responsible for the failure. In that case, the solution message had no indirect effect on satisfaction (via competence), while the empathy message had a positive indirect effect on satisfaction. When managers are in doubt about whether the chatbot or user is responsible for the failure, the empathy message reflects a rather uncritical choice. In sum, our results show that the ‘solution’ message is more effective than the ‘empathy’ message in some situations, while it is the other way round in other situations. Therefore, managers need to be aware of the type of failure to evaluate failure attributions, and about the failure frequency, in order to adapt the recovery messages accordingly.

More generally, with the fast-paced developments in the area of deep learning and large language models, managers might be tempted to integrate chatbots such as “ChatGPT” in their service processes (Dwivedi et al., 2023). However, unlike most current chatbots (based on natural language processing or simple decision trees), which respond generally with some sort of error message (e.g., "Sorry, I don't know"), ChatGPT generally responds with a text expressing the most likely answer. Based on a vast amount of available text, the algorithm aims to anticipate the highest likelihood of an answer by forecasting what a human would use to reply to the specific request. Thus, instead of acknowledging failure, ChatGPT often “hallucinates”, meaning that these kinds of chat tools produce information that may be nonsensical, untrue, or inconsistent with the content of the source input (Dwivedi et al., 2023; Ji et al., 2023). In the context of diverse service interactions, such hallucinated responses to user queries pose a significant threat, as service activities are often associated with actions (e.g., customer data, confirmations, bookings, returns, etc.). Therefore, while the integration of language processing models such as ChatGPT may be beneficial for service interactions, failure acknowledgement and recovery attempts (e.g., via messages) remain highly relevant for digital service interactions.

Limitations and future research

Although this research offers valuable insights, it also has some limitations.

First, our study relied on screenshots of chat conversations to ensure high internal validity. Future research could apply the results in a field study with real-life chats on firm websites or messenger services to increase external validity. In this vein, scholars could also analyze if new and more sophisticated bots such as ChatGPT are less prone to service failures, and whether these bots could also integrate more context-aware information to create a more personalized and failure-congruent recovery message.

Second, this study considered failure frequency and failure attribution as two situational factors. Future studies could include additional factors such as different message tonalities or recovery feedback elements (e.g., offering a message and asking if the information was helpful). Related chatbot studies revealed that already minor adaptations in the conversational design (e.g., response delays or chatbot- vs. user-initiation) may have effects on user's satisfaction with the chatbot (Gnewuch et al., 2018; Pizzi et al., 2021). Thus, future research might explore combinations of such design elements with message content elements.

Third, while our research did not focus on the role of emotions in chatbot failure and recovery, prior research found emotions to influence consumers' reactions in chatbot interactions (Crolic et al., 2021). Future studies should therefore investigate the role of emotions such as anger, frustration and helplessness in human-chatbot interactions.

Fourth, we used two prototypical messages to measure their effects precisely, neglecting other possible forms or mixtures of messages, or even the combination with other forms of compensation such as vouchers or human interaction, leaving open a fruitful field for future research related to digital agents' conversational design.

5. References

Aaker, J., Vohs, K. D., & Mogilner, C. (2010). Nonprofits Are Seen as Warm and For-Profits as Competent: Firm Stereotypes Matter. *Journal of Consumer Research*, *37*(2), 224–237. https://doi.org/10.1086/651566

Adam, M., Wessel, M., & Benlian, A. (2020). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets.* Advance online publication. https://doi.org/10.1007/s12525-020-00414-7

Agustin, C., & Singh, J. (2005). Curvilinear Effects of Consumer Loyalty Determinants in Relational Exchanges. *Journal of Marketing Research*, *42*(1), 96–108. https://doi.org/10.1509/jmkr.42.1.96.56961

Akhtar, M., Neidhardt, J., & Werthner, H. (2019). The Potential of Chatbots: Analysis of Chatbot Conversations. In M. Akhtar, J. Neidhardt, & H. Werthner (Eds.), *Potential of Chatbots: Analysis of Chatbot Conversations* (pp. 397–404). IEEE. https://doi.org/10.1109/CBI.2019.00052

Alt, R. (2020). Evolution and perspectives of electronic markets. *Electronic Markets*, *30*(1), 1–13. https://doi.org/10.1007/s12525-020-00413-8

Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient Chatbots. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (pp. 1–12). ACM Press. https://doi.org/10.1145/3290605.3300484

Babbar, S., & Koufteros, X. (2008). The human element in airline service quality: contact personnel and the customer. *International Journal of Operations & Production Management*, *28*(9), 804–830. https://doi.org/10.1108/01443570810895267

Belanche, D., Casaló, L. V., Schepers, J., & Flavián, C. (2021). Examining the effects of robots' physical appearance, warmth, and competence in frontline services: The Humanness-Value-Loyalty model. *Psychology & Marketing*, *38*(12), 2357–2376. https://doi.org/10.1002/mar.21532

Benner, D., Elshan, E., Schöbel, S., & Janson, A. (2021). What do you mean? A Review on Recovery Strategies to Overcome Conversational Breakdowns of Conversational Agents. *Proceedings of the 42. International Conference on Information Systems (ICIS)*, 1–17.

Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots:: changing user needs and motivations. *Interactions*, *25*(5), 38–43. https://doi.org/10.1145/3236669

Brendel, A. B., Greve, M., Diederich, S., Bührke, J., & Kolbe, L. M. (2020). 'You are an Idiot!' – How Conversational Agent Communication Patterns Influence Frustration and Harassment". *Proceedings of Americas Conference on Information Systems*.

Choi, S [Sungwoo], Mattila, A. S., & Bolton, L. E. (2021). To Err Is Human(-oid): How Do Consumers React to Robot Service Failure and Recovery? *Journal of Service Research*, *24*(3), 354–371. https://doi.org/10.1177/1094670520978798

Choi, S [Sunmee], & Mattila, A. S. (2008). Perceived controllability and service expectations: Influences on customer reactions following service failure. *Journal of Business Research*, *61*(1), 24–30. https://doi.org/10.1016/j.jbusres.2006.05.006

Chong, T., Yu, T [Ting], Keeling, D. I., & Ruyter, K. de (2021). AI-chatbots on the services frontline addressing the challenges and opportunities of agency. *Journal of Retailing and Consumer Services*, *63*, 102735. https://doi.org/10.1016/j.jretconser.2021.102735

Collier, J. E., Breazeale, M., & White, A. (2017). Giving back the "self" in self service: customer preferences in self-service failure recovery. *Journal of Services Marketing*, *31*(6), 604–617. https://doi.org/10.1108/JSM-07-2016-0259

Crolic, C., Thomaz, F., Hadi, R., & Stephen, A. T. (2021). Blame the Bot: Anthropomorphism and Anger in Customer-Chatbot Interactions. *Journal of Marketing*, 002224292110456. https://doi.org/10.1177/00222429211045687

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology. Advances in experimental social psychology* (Vol. 40, pp. 61–149). Elsevier. https://doi.org/10.1016/S0065-2601(07)00002-0

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2020). Designing Anthropomorphic Enterprise Conversational Agents. *Business & Information Systems Engineering*, *62*(3), 193–209. https://doi.org/10.1007/s12599-020-00639-y

Dilmegani, C. (2022, December 21). *Chatbot: 9 Epic Chatbot/Conversational Bot Failures*. https://research.aimultiple.com/chatbot-fail/

Dubois, D., Rucker, D. D., & Galinsky, A. D. (2016). Dynamics of Communicator and Audience Power: The Persuasiveness of Competence versus Warmth. *Journal of Consumer Research*, *43*(1), 68–85. https://doi.org/10.1093/jcr/ucw006

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Gelbrich, K. (2010). Anger, frustration, and helplessness after service failure: coping strategies and effective informational support. *Journal of the Academy of Marketing Science*, *38*(5), 567–585. https://doi.org/10.1007/s11747-009-0169-6

Gerdes, K. E. (2011). Empathy, Sympathy, and Pity: 21st-Century Definitions and Implications for Practice and Research. *Journal of Social Service Research*, *37*(3), 230–241. https://doi.org/10.1080/01488376.2011.564027

Gnewuch, U., Morana, S., Adam, M. T., & & Maedche, A. (2018). Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. *European Conference on Information Systems*. https://www.semanticscholar.org/paper/Faster-is-Not-Always-Better%3A-Understanding-the-of-Gnewuch-Morana/22cbf658ea99b2901b3f6f649e21ef8a3c7a590d

Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards Designing Cooperative and Social Conversational Agents for Customer Service. *Proceedings of the International Conference on Information Systems*, *38*.

Güntürkün, P., Haumann, T., & Mikolon, S. (2020). Disentangling the Differential Roles of Warmth and Competence Judgments in Customer-Service Provider Relationships.

Journal of Service Research, *23*(4), 476–503. https://doi.org/10.1177/1094670520920354

Habel, J., Alavi, S., & Pick, D. (2017). When serving customers includes correcting them: Understanding the ambivalent effects of enforcing service rules. *International Journal of Research in Marketing*, *34*(4), 919–941. https://doi.org/10.1016/j.ijresmar.2017.09.002

Hayes, A. F. (2018). *Introduction to Mediation, Moderation and Conditional Process Analysis. A Regression-Based Approach* (2nd edition). The Guilford Press.

Hein, A., Schreieck, M., Riasanow, T., Setzke, D. S., Wiesche, M., Böhm, M., & Krcmar, H. (2020). Digital platform ecosystems. *Electronic Markets*, *30*(1), 87–98. https://doi.org/10.1007/s12525-019-00377-4

Hogan, R. (1969). Development of an empathy scale. *Journal of Consulting and Clinical Psychology*, *33*(3), 307–316. https://doi.org/10.1037/h0027580

Hosseini, M., & Caragea, C. (2021). It Takes Two to Empathize: One to Seek and One to Provide. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(14), 13018–13026.

Huang, R., & Ha, S. (2020). The effects of warmth-oriented and competence-oriented service recovery messages on observers on online platforms. *Journal of Business Research*, *121*, 616–627. https://doi.org/10.1016/j.jbusres.2020.04.034

Hulland, J., Baumgartner, H., & Smith, K. M. (2018). Marketing survey research best practices: evidence and recommendations from a review of JAMS articles. *Journal of the Academy of Marketing Science*, *46*(1), 92–108. https://doi.org/10.1007/s11747-017-0532-y

Janssen, A., Grützner, L., & Breitner, M. H. (2021). Why do Chatbots fail? A Critical Success Factors Analysis. *Proceedings of the International Conference on Information Systems (ICIS)*.

Ji, Z., Lee, N., Frieske, R., Yu, T [Tiezheng], Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Johnston, R., & Fern, A. (1999). Service Recovery Strategies for Single and Double Deviation Scenarios. *The Service Industries Journal*, *19*(2), 69–82. https://doi.org/10.1080/02642069900000019

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology*, *89*(6), 899–913. https://doi.org/10.1037/0022-3514.89.6.899

Kull, A. J., Romero, M., & Monahan, L. (2021). How may I help you? Driving brand engagement through the warmth of an initial chatbot message. *Journal of Business Research*, *135*, 840–850. https://doi.org/10.1016/j.jbusres.2021.03.005

Lazarus, R. S. (1991). *Emotion and Adaptation.* Oxford University Press.

Li, C [Charlene], Mirosa, M., & Bremer, P. (2020). Review of Online Food Delivery Platforms and their Impacts on Sustainability. *Sustainability*, *12*(14), 5528. https://doi.org/10.3390/su12145528

Luger, E., & Sellen, A. (2016). "Like Having a Really Bad PA". In J. Kaye, A. Druin, C. Lampe, D. Morris, & J. P. Hourcade (Eds.), *#chi4good: Chi 2016 : San Jose, CA, USA, May 7-12 : Proceedings : The 34th Annual CHI Conference on Human Factors* *in Computing Systems : San Jose Convention Center* (pp. 5286–5297). The Association for Computing Machinery. https://doi.org/10.1145/2858036.2858288

Lv, L., Huang, M., Guan, D., & Yang, K. (2022). Apology or gratitude? The effect of communication recovery strategies for service failures of AI devices. *Journal of Travel & Tourism Marketing*, *39*(6), 570–587. https://doi.org/10.1080/10548408.2022.2162659

Lv, X., Liu, Y [Yue], Luo, J., Liu, Y [Yuqing], & Li, C [Chunxiao] (2021). Does a cute artificial intelligence assistant soften the blow? The impact of cuteness on customer tolerance of assistant service failure. *Annals of Tourism Research*, *87*, 103114. https://doi.org/10.1016/j.annals.2020.103114

Lv, X., Yang, Y., Qin, D., Cao, X., & Xu, H. (2022). Artificial intelligence service recovery: The role of empathic response in hospitality customers' continuous usage intention. *Computers in Human Behavior*, *126*, 106993. https://doi.org/10.1016/j.chb.2021.106993

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants. *Business & Information Systems Engineering*, *61*(4), 535–544. https://doi.org/10.1007/s12599-019-00600-8

Marinova, D., Singh, S. K., & Singh, J. (2018). Frontline Problem-Solving Effectiveness: A Dynamic Analysis of Verbal and Nonverbal Cues. *Journal of Marketing Research*, *55*(2), 178–192. https://doi.org/10.1509/jmr.15.0243

McKee, K. R., Bai, X., & Fiske, S. T. (2022). Warmth and competence in human-agent cooperation. *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022).*

Misselhorn, C. (2009). Empathy with Inanimate Objects and the Uncanny Valley. *Minds and Machines*, *19*(3), 345–359. https://doi.org/10.1007/s11023-009-9158-2

Mordor Intelligence. (2021). *Chatbot Market: Growth, Trends, COVID-19 Impact, and Forecasts*. https://www.mordorintelligence.com/industry-reports/chatbot-market

Mostafa, R. B., & Kasamani, T. (2021). Antecedents and consequences of chatbot initial trust. *European Journal of Marketing.* Advance online publication. https://doi.org/10.1108/EJM-02-2020-0084

Mozafari, N., Weiger, W. H., & Hammerschmidt, M. (2021). That's so Embarrassing! When not to Design for Social Presence in Human–Chatbot. *ICIS 2021 Proceedings*, *5*. https://aisel.aisnet.org/icis2021/hci_robot/hci_robot/5

Mozafari, N., Weiger, W. H., & Hammerschmidt, M. (2022). Trust me, I'm a bot – repercussions of chatbot disclosure in different service frontline settings. *Journal of Service Management*, *33*(2), 221–245. https://doi.org/10.1108/JOSM-10-2020-0380

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, *45*(6), 669–678. https://doi.org/10.1006/ijhc.1996.0073

Olson, P. (2022). *Google Faces a Serious Threat From ChatGPT.* https://www.washingtonpost.com/business/energy/google-faces-a-serious-threat-from-chatgpt/2022/12/07/363d2440-75f5-11ed-a199-927b334b939f_story.html

Ozgen, O., & Duman Kurt, S. (2012). Pre-recovery and post-recovery emotions in the service context: a preliminary study. *Managing Service Quality: An International Journal*, *22*(6), 592–605. https://doi.org/10.1108/09604521211287561

Pacheco, N. A., Pizzutti, C., Basso, K., & van Vaerenbergh, Y. (2019). Trust recovery tactics after double deviation: better sooner than later? *Journal of Service Management*, *30*(1), 2–22. https://doi.org/10.1108/JOSM-02-2017-0056

Pizzi, G., Scarpi, D., & Pantano, E. (2021). Artificial intelligence and the new forms of interaction: Who has the control when interacting with a chatbot? *Journal of Business Research*, *129*, 878–890. https://doi.org/10.1016/j.jbusres.2020.11.006

Poser, M., Singh, S., & Bittner, E. (2021). Hybrid Service Recovery: Design for Seamless Inquiry Handovers between Conversational Agents and Human Service Agents. In T. Bui (Ed.), *Proceedings of the Annual Hawaii International Conference on System Sciences, Proceedings of the 54th Hawaii International Conference on System Sciences.* Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.144

Qualtrix and ServiceNow. (2021). *Customer Experience*. https://www.qualtrics.com/blog/qualtrics-servicenow-customer-service-research/

Reinkemeier, F., & Gnewuch, U. (2022). Designing Effective Conversational Repair Strategies for Chatbots. *European Conference on Information Systems (ECIS), Timișoara, Romania*, *30*.

Roschk, H., & Gelbrich, K. (2014). Identifying Appropriate Compensation Types for Service Failures. *Journal of Service Research*, *17*(2), 195–211. https://doi.org/10.1177/1094670513507486

Sands, S., Campbell, C., Plangger, K., & Pitt, L. (2022). Buffer bots: The role of virtual service agents in mitigating negative effects when service fails. *Psychology & Marketing*, *39*(11), 2039–2054. https://doi.org/10.1002/mar.21723

Sands, S., Ferraro, C., Campbell, C., & Tsao, H.-Y. (2021). Managing the human–chatbot divide: how service scripts influence service experience. *Journal of Service Management*, *32*(2), 246–264. https://doi.org/10.1108/JOSM-06-2019-0203

Scott, M. L., Mende, M., & Bolton, L. E. (2013). Judging the Book by Its Cover? How Consumers Decode Conspicuous Consumption Cues in Buyer–Seller Relationships. *Journal of Marketing Research*, *50*(3), 334–347. https://doi.org/10.1509/jmr.11.0478

Seeger, A.-M., & Heinzl, A. (2021). Chatbots often Fail! Can Anthropomorphic Design Mitigate Trust Loss in Conversational Agents for Customer Service? *Proceedings of the European Conference on Information Systems*, *12*.

Sheehan, B., Jin, H. S., & Gottlieb, U. (2020). Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research*, *115*, 14–24. https://doi.org/10.1016/j.jbusres.2020.04.030

Song, M., Du, J., Xing, X., & Mou, J. (2022). Should the chatbot “save itself” or “be helped by others”? The influence of service recovery types on consumer perceptions of recovery satisfaction. *Electronic Commerce Research and Applications*, *55*, 101199. https://doi.org/10.1016/j.elerap.2022.101199

Song, M., Zhang, H., Xing, X., & Duan, Y [Yucong] (2023). Appreciation vs. apology: Research on the influence mechanism of chatbot service recovery based on politeness theory. *Journal of Retailing and Consumer Services*, *73*, 103323. https://doi.org/10.1016/j.jretconser.2023.103323

Sporrer, T. (2021). *Your Chatbot Is Frustrating Your Customers: Improve the Experience with Guidance*. https://customerthink.com/your-chatbot-is-frustrating-your-customers-improve-the-experience-with-guidance/

Stoeckli, E., Dremel, C., Uebernickel, F., & Brenner, W. (2020). How affordances of chatbots cross the chasm between social and traditional enterprise systems. *Electronic Markets*, *30*(2), 369–403. https://doi.org/10.1007/s12525-019-00359-6

Suta, P., Lan, X., Wu, B., Mongkolnam, P., & Chan, J. H. (2020). An Overview of Machine Learning in Chatbots. *International Journal of Mechanical Engineering and Robotics Research*, 502–510. https://doi.org/10.18178/ijmerr.9.4.502-510

Thorbecke, C. (2022). *Chatbots: A long and complicated history.* CNN. https://edition.cnn.com/2022/08/20/tech/chatbot-ai-history/index.html

van der Goot, M. J., Hafkamp, L., & Dankfort, Z. (2021). Customer Service Chatbots: A Qualitative Interview Study into Customers' Communication Journey. In A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, E. Luger, M. Goodwin, & P. B. Brandtzaeg (Eds.), *Chatbot Research and Design* (pp. 190–204). Springer International Publishing.

van Doorn, J., Mende, M., Noble, S. M., Hulland, J., Ostrom, A. L., Grewal, D., & Petersen, J. A. (2016). Domo Arigato Mr. Roboto. *Journal of Service Research*, *20*(1), 43–58. https://doi.org/10.1177/1094670516679272

van Pinxteren, M. M., Pluymaekers, M., & Lemmink, J. G. (2020). Human-like communication in conversational agents: a literature review and research agenda. *Journal of Service Management*, *31*(2), 203–225. https://doi.org/10.1108/JOSM-06-2019-0175

van Vaerenbergh, Y., Varga, D., Keyser, A. de, & Orsingher, C. (2019). The Service Recovery Journey: Conceptualization, Integration, and Directions for Future Research. *Journal of Service Research*, *22*(2), 103–119. https://doi.org/10.1177/1094670518819852

Wallenburg, C. M., & Lukassen, P. (2011). Proactive improvement of logistics service providers as driver of customer loyalty. *European Journal of Marketing*, *45*(3), 438–454. https://doi.org/10.1108/03090561111107267

Weiler, S., Matt, C., & Hess, T. (2022). Immunizing with information - Inoculation messages against conversational agents' response failures. *Electronic Markets*, *32*(1), 239–258. https://doi.org/10.1007/s12525-021-00509-9

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548–573. https://doi.org/10.1037/0033-295X.92.4.548

Weiner, B. (2012). An attribution theory of motivation. In P. A. VanLange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (1st ed., pp. 135–155). Sage.

Wieseke, J., Geigenmüller, A., & Kraus, F. (2012). On the Role of Empathy in Customer-Employee Interactions. *Journal of Service Research*, *15*(3), 316–331. https://doi.org/10.1177/1094670512439743

Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., Lu, V. N., Paluch, S., & Martins, A. (2018). Brave new world: service robots in the frontline. *Journal of Service Management*, *29*(5), 907–931. https://doi.org/10.1108/JOSM-04-2018-0119

Worsfold, K., Worsfold, J., & Bradley, G. (2007). Interactive Effects of Proactive and Reactive Service Recovery Strategies: The Case of Rapport and Compensation. *Journal of Applied Social Psychology*, *37*(11), 2496–2517. https://doi.org/10.1111/J.1559-1816.2007.00267.X

Xie, C., Bagozzi, R. P., & Grønhaug, K. (2015). The role of moral emotions and individual differences in consumer responses to corporate green and non-green actions. *Journal* *of the Academy of Marketing Science*, *43*(3), 333–356. https://doi.org/10.1007/s11747-014-0394-5

Xing, X., Song, M., Duan, Y [Yucong], & Mou, J. (2022). Effects of different service failure types and recovery strategies on the consumer response mechanism of chatbots. *Technology in Society*, *70*, 102049. https://doi.org/10.1016/j.techsoc.2022.102049

Xu, X., & Liu, J. (2022). Artificial intelligence humor in service recovery. *Annals of Tourism Research*, *95*, 103439. https://doi.org/10.1016/j.annals.2022.103439

Yang, Z., Zhou, J., & Yang, H. (2023). The Impact of AI's Response Method on Service Recovery Satisfaction in the Context of Service Failure. *Sustainability*, *15*(4), 3294. https://doi.org/10.3390/su15043294

You, Y., Yang, X., Wang, L., & Deng, X. (2020). When and Why Saying "Thank You" Is Better Than Saying "Sorry" in Redressing Service Failures: The Role of Self-Esteem. *Journal of Marketing*, *84*(2), 133–150. https://doi.org/10.1177/0022242919889894

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1996). The Behavioral Consequences of Service Quality. *Journal of Marketing*, *60*(2), 31–46. https://doi.org/10.2307/1251929

Figure 1. Conceptual model

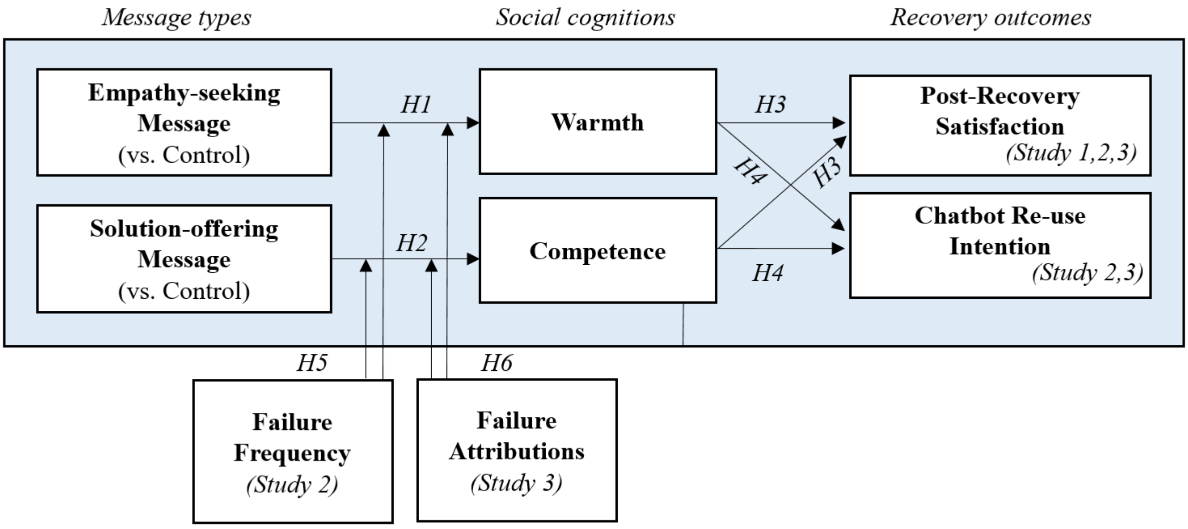
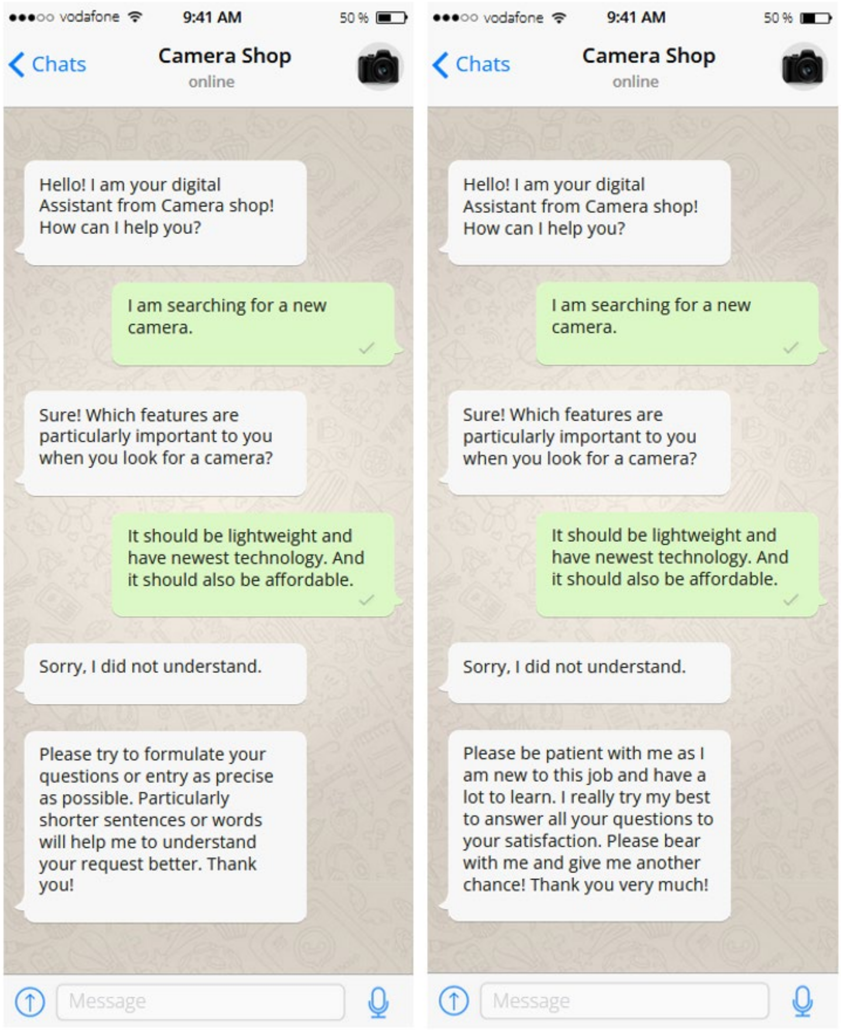

Message types
Social cognitions
Recovery outcomes
Empathy-seeking Message (vs. Control)
Solution-offering Message (vs. Control)
H1
H2
Warmth
Competence
H3
H4
H4
H3
Post-Recovery Satisfaction (Study 1,2,3)
Chatbot Re-use Intention (Study 2,3)
H5
H6
Failure Frequency (Study 2)
Failure Attributions (Study 3)

Figure 2. Exemplary scenarios



Hello! I am your digital Assistant from Camera shop! How can I help you?

I am searching for a new camera.

Sure! Which features are particularly important to you when you look for a camera?

It should be lightweight and have newest technology. And it should also be affordable.

Sorry, I did not understand.

Please try to formulate your questions or entry as precise as possible. Particularly shorter sentences or words will help me to understand your request better. Thank you!

Solution message

Hello! I am your digital Assistant from Camera shop! How can I help you?

I am searching for a new camera.

Sure! Which features are particularly important to you when you look for a camera?

It should be lightweight and have newest technology. And it should also be affordable.

Sorry, I did not understand.

Please be patient with me as I am new to this job and have a lot to learn. I really try my best to answer all your questions to your satisfaction. Please bear with me and give me another chance! Thank you very much!

Empathy message

Table 1. Literature Review – Chatbots Service Failure and Recovery

Research Streams	**Source**	**Topic**	**Literature Field(s)**	**Mediator(s), Moderator(s)**	**Main Findings**
Conceptual / Qualitative	Benner et al., 2021	Literature review on chatbot failure recovery strategies	IS, HCI	-	Literature analysis presents six categories of chatbots' conversational breakdown recovery strategies.
	Janssen et al., 2021	Reasons of chatbot failures and critical factors for success	Design science, IS, HCI		Based on literature review and expert interviews, major reasons for chatbot failures and 12 critical success factors are identified.
	Poser et al., 2021	Design principles for handover from chatbot to human agent after failure	Design science, HCI	-	Experts confirmed positive perception of real-time hand-over to human agent when chatbots fail to resolve requests. The process of instant hand-over is proposed. Experts confirmed a positive perception of this hybrid approach for service recovery.
Pre-failure interaction strategies (to mitigate failure severity perception)	Sands et al., 2022	Effects of agent type (human vs. chatbot) related to failure type and magnitude	Customer service, IS	Mod: Service failure magnitude	Consumers respond less negative to a process failure when it occurs in human agent interaction (vs. chatbot as virtual agent). In case of a large outcome failure, consumers respond more positively when the agent is virtual (vs. human).
	Seeger and Heinzl, 2021	Anthropomorphic design to protect trust levels from damages due to service failure	IS, HCI	Med: Perceived norm violation	Anthropomorphic design of a chatbot mitigates negative effect of a failure on trust and Word-of-mouth.
	Sheehan et al., 2020	Effects of clarification technique on anthropomorphism	HCI, customer service	Med: Anthropomorphism Mod: Need for human interaction	Chatbot interactions using clarification technique are similarly effective regarding chatbot adoption compared to a perfectly understanding (i.e. error-free) chatbot, and both types are significantly more effective than a chatbot which failed to understand a user input. These effects are

Pre-failure interaction strategies (continued)		and adoption intention			mediated by higher anthropomorphism compared to a failing chatbot.
	Weiler et al., 2021	Effects of ex-ante messages to reduce discontinuance after chatbot failures	Customer service	Mod: Performance level (low vs. high), Linguistic form (quantitative vs. qualitative)	Inoculation messages (i.e., preparing users for possible failures) reduce user's decision to discontinue chatbot after service failure. Quantitative form of the message moderated the effect of perceived performance on discontinuance, whereas a qualitative form moderated the effect of trusting beliefs.
Post-failure recovery strategies	Ashktorab et al., 2019	User preferences of different chatbot failure repair strategies	HCI, Conversational design	-	User preferences of eight chatbot notification strategies after a chatbot response failure. Notifications which provide options or explanations are favored as they reflect a chatbot's initiative and are actionable to recover from failures.
	Brendel et al., 2020	Effects of three communication patterns on user's emotions and responses	IS, HCI	-	After a service failure, a "helping" communication style user leads to low levels of harassment, whereas both a "persecutor" and a "victim" style lead to significantly higher harassment. User frustration is not different across communication styles, but post-failure satisfaction is significantly higher for "helping" and "victim" (vs. "persecutor").
	Lv (Linxiang) et al., 2022	Effects of communication recovery strategies for AI device failures		Med: Relational needs and efficacy needs.	When chatbot message expresses gratitude (vs. apology), consumers are more likely to forgive a failure if service failures includes being rejected (vs. being ignored).
	Lv (Xingyang) et al., 2021	Effects of cuteness after AI service failure	Design science, customer service	Med: tenderness and performance expectancy Mod: failure severity, time pressure	Cuteness increases consumers' tolerance of AI service failure, mediated by tenderness and performance expectancy.

Post-failure recovery strategies (continued)	Lv (Xingyang) et al., 2022	Effects of empathic response on users continuance intentions	HCI, Customer service	Med: Psychological distance, trust	Highly empathic AI response increases continuance intentions. Multisensory (i.e., text and audio) interactions amplify the effect, compared to text-only interactions.
	Mozafari et al., 2022	Effects of chatbot disclosure depending on service-related factors	Service	Med: Trust Mod: Service criticality, Failure vs. no failure	In case of service failure, chatbot identity disclosure leads to positive effect on retention. Without a failure, disclosure negatively affects customer retention through reduced trust for services with high levels of criticality.
	Song et al., 2022	Effect of chatbot self-recovery vs. human support on recovery satisfaction	IS, Customer service	Mod: Chatbot intelligence	Chatbot self-recovery has a more positive effect on post-recovery satisfaction compared to human-supported recovery. Effect is moderated by chatbot intelligence
	Song et al., 2023	Effects of two politeness strategies on post-recovery satisfaction	Customer service	Med: Face concern, Mod: Time pressure	Focusing on human-chatbot relationship with an appreciation message led to more post-recovery satisfaction compared to an admittance of the limited abilities with an apology message.
	Xing et al., 2022	Effects of service failure types on consumers' recovery strategy choices	HCI, conversational design	Mod: Chatbot intelligence	After a functional (nonfunctional) failure, consumers tend to use the chatbot (a human) for service recovery. Chatbot intelligence moderates this choice. Chatbots (vs. human employees) have higher perceived governance.
	Xu et al., 2022	Effects of humor in messages after AI service failure	HCI, Conversational design	Med: Perceived warmth and competence Mod: Time pressure, customer inoculation	Humorous response of chatbot after service failure increases consumer tolerance, mediated by warmth and competence.

Post-failure recovery strategies (continued)	Yang et al., 2023	Impact of AI’s self-deprecating humor on service recovery satisfaction	Customer service, conversational design	Med: Perceived intelligence, perceived sincerity Mod: Failure experience, Sense of power	Self-deprecating humor improves consumers’ willingness to tolerate failure and recovery satisfaction, mediated by higher perceptions of intelligence and sincerity. Effect of humor was present in process-related failures, but not in outcome-related failures.
	Our Study	Effectiveness of two failure-recovery message strategies depending on situational factors	Customer service, Conversational design	Med: Perceived warmth and competence	Using an empathy-seeking (solution-suggesting) message as recovery strategy increases post-recovery satisfaction and re-use intention, mediated by warmth (competence), respectively. Moreover, message preference depends on failure attribution (user vs. chatbot) and failure frequency.

Table 2. Scenario descriptions

Solution Message Scenarios	**Empathy Message Scenarios**	**Control**
Study 1 *(Electronics provider)*	***Study 1*** *(Electronics provider)*	***Study 1*** *(Electronics provider)*
[Chatbot]: Hello! I am your digital Assistant from [Camera shop]! How can I help you? [User:] I am searching for a new camera. [Chatbot]: Sure! Which features are particularly important to you when you look for a camera? [User:] It should be lightweight and have newest technology. And it should also be affordable. [Chatbot]: Sorry, I did not understand. [Chatbot]: Please try to formulate your questions or entry as precise as possible. Particularly shorter sentences or words will help me to understand your request better. Thank you!	[Chatbot]: Hello! I am your digital Assistant from [Camera shop]! How can I help you? [User:] I am searching for a new camera. [Chatbot]: Sure! Which features are particularly important to you when you look for a camera? [User:] It should be lightweight and have newest technology. And it should also be affordable. [Chatbot]: Sorry, I did not understand. [Chatbot]: Please be patient with me as I am new to this job and have a lot to learn. I really try my best to answer all your questions to your satisfaction. Please bear with me and give me another chance! Thank you very much!	[Chatbot]: Hello! I am your digital Assistant from [Camera shop]! How can I help you? [User:] I am searching for a new camera. [Chatbot]: Sure! Which features are particularly important to you when you look for a camera? [User:] It should be lightweight and have newest technology. And it should also be affordable. [Chatbot]: Sorry, I did not understand.
Study 2 *(Table reservation)*	***Study 2*** *(Table reservation)*	***Study 2*** *(Table reservation)*
[Chatbot]: Hello! I am your digital Assistant from Pizza House! How can I help you? [User:] I would like to reserve a table for next week	[Chatbot]: Hello! I am your digital Assistant from Pizza House! How can I help you? [User:] I would like to reserve a table for next week Tuesday between 12 and 13h for me and my partner.	[Chatbot]: Hello! I am your digital Assistant from Pizza House! How can I help you? [User:] I would like to reserve a table for next week Tuesday between 12 and 13h for me and my partner. [Chatbot]: Sorry, I did not understand.

Tuesday between 12 and 13h for me and my partner. [Chatbot]: Sorry, I did not understand. [Chatbot]: Please try to formulate your questions or entry as precise as possible. Particularly shorter sentences or precise words will help me to understand your request better. Thank you! [User]: [individual entry] [Chatbot]: Sorry, I did not understand you again. Please formulate your entry as precise as possible. For example, please enter only the date, arrival time and the number of people for the reservation. [2nd failure] / Thank you! I booked a table for Tuesday for flexible arrival between 12:00 and 13:00 for 2 people. Your booking code is 2553. See you soon! [Recovery success]	[Chatbot]: Sorry, I did not understand. [Chatbot]: Please be patient with me as I am new to this job and have a lot to learn. I really try my best to answer all your questions to your satisfaction. Please bear with me and give me another chance. Thank you very much! [User]: [individual entry] [Chatbot]: Sorry, I did not understand you again. Please be patient with me again as I am new to this job and have a lot to learn. I really try my best to answer all your questions to your satisfaction. Please bear with me and give me another chance. Thank you very much! [2nd failure] / Thank you! I booked a table for Tuesday for flexible arrival between 12:00 and 13:00 for 2 people. Your booking code is 2553. See you soon! [Recovery success]	[User]: [individual entry] [Chatbot]: Sorry, I did not understand you again. [2nd failure] / Thank you! I booked a table for Tuesday for flexible arrival between 12:00 and 13:00 for 2 people. Your booking code is 2553. See you soon! [Recovery success]
Study 3 *(Pizza delivery)* [Chatbot]: Hello! I am your digital Assistant from Pizza House! How can I help you? [User:] I would like to order a Pizza Salami to my home [Chatbot]: Sure! To which address may I deliver your Pizza? [User:] To Schlösschen Street 12, please [Chatbot-attributed failure] / "To my	***Study 3*** *(Pizza delivery)* [Chatbot]: Hello! I am your digital Assistant from Pizza House! How can I help you? [User:] I would like to order a Pizza Salami to my home [Chatbot]: Sure! To which address may I deliver your Pizza? [User:] To Schlösschen Street 12, please [Chatbot-attributed failure] / "To my	***Study 3*** *(Pizza delivery)* [Chatbot]: Hello! I am your digital Assistant from Pizza House! How can I help you? [User:] I would like to order a Pizza Salami to my home [Chatbot]: Sure! To which address may I deliver your Pizza? [User:] To Schlösschen Street 12, please [Chatbot-attributed failure] / "To my

home please” [User-attributed failure] [Chatbot]: Sorry, I did not understand. I do not know the address "To Schlösschen Street 12 please" [Chatbot-attributed failure] / “To my home please” [User-attributed failure]. [Chatbot]: Please try to formulate your questions as precise as possible. Particularly shorter sentences or words will help me to understand your request better. Thank you!	home please” [User-attributed failure] [Chatbot]: Sorry, I did not understand. I do not know the address "To Schlösschen Street 12 please" [Chatbot-attributed failure] / “To my home please” [User-attributed failure]. [Chatbot]: Please be patient with me as I am new to this job and have a lot to learn. I really try my best to answer all your questions to your satisfaction. Please bear with me and give me another chance! Thank you very much!	home please” [User-attributed failure] [Chatbot]: Sorry, I did not understand.

Table 3. Scale items and statistics

Construct name and items	Factor loading		
	Study 1 -	Study 2a/b -	Study 3
Warmth (Study 1/2a and b/3: α = .80/.95 and .96/.87; CR = .87/ .92 and .93/ .91; AVE = .69 / .80 and .81/. 76)			
I perceive the chatbot as …			
• warm	.80	.90/.91	.84
• kind	.84	.91/.92	.88
• generous	.85	.87/.87	.91
Competence (Study 1/2a and b/3: α = .87/.95 and .98/ .90; CR = .88 /.90 and .89/ .88; AVE = .71 /.75 and .72/ .71)			
I perceive the chatbot as …			
• competent	.77	.84/.86	.77
• effective	.87	.88/.84	.90
• efficient	.89	.88/.85	.86
Post-recovery Satisfaction (Study 1/2a and b/3: α = .74/.93 and .98/.90; CR = .80 /.83 and .86/ .82; AVE = .58 /.62 and .67/ .60)			
The interaction with the chatbot service was …			
• satisfying	.79	.74/.84	.73
• pleasant	.64	.81/.79	.81
• good	.83	.81/.83	.79
Re-use Intentions (Study 2b/3: α = .96/.96, CR = .85/.87; AVE = .65/.70)			
• I would use this chatbot again		- /.79	.79
• I would use this chat service in my daily life		- /.83	.89
• I would order my pizza again with this chatbot		- /.80	.83

Note. Study 2a: values after first failure; 2b: values after second response

Table 4. Descriptives for Study 1, 2 and 3

	Dependent Variables			
	Warmth	Competence	Post-recovery Satisfaction	Re-use Intentions
Study 1				
Empathy	4.40 (1.43)	2.66 (1.27)	2.11 (1.03)	-
Solution	3.87 (1.30)	3.19 (1.29)	2.50 (1.19)	-
Control	3.53 (1.33)	2.66 (1.35)	1.78 (0.79)	-
Study 2				
First failure				
Empathy	4.35 (1.51)	2.88 (1.61)	3.08 (1.48)	-
Solution	3.64 (1.49)	3.51 (1.52)	3.18 (1.56)	-
Control	3.04 (1.61)	2.78 (1.52)	2.41 (1.39)	-
Second failure				
Empathy	3.59 (1.64)	1.80 (1.11)	1.95 (1.14)	1.86 (1.25)
Solution	3.28 (1.54)	2.66 (1.75)	2.66 (1.79)	2.23 (1.87)
Control	2.57 (1.49)	1.90 (1.32)	1.77 (1.18)	2.17 (1.34)
Success				
Empathy	4.67 (1.73)	5.17 (1.50)	5.10 (1.65)	4.63 (1.81)
Solution	3.69 (1.55)	5.02 (1.03)	4.64 (1.22)	4.81 (1.36)
Control	3.85 (1.84)	4.76 (1.61)	4.81 (1.61)	4.68 (1.69)
Study 3				
User-attributed failure				
Empathy	4.63 (1.41)	3.66 (1.47)	4.49 (1.62)	4.37 (1.89)
Solution	3.70 (1.16)	4.54 (1.49)	4.31 (1.53)	4.13 (1.70)
Control	3.48 (1.36)	3.79 (1.34)	4.27 (1.65)	4.02 (1.78)
Chatbot-attributed failure				
Empathy	4.28 (1.30)	3.21 (1.15)	3.63 (1.22)	3.33 (1.70)
Solution	3.53 (1.53)	3.49 (1.72)	3.15 (1.39)	3.46 (1.99)
Control	3.06 (1.35)	2.75 (1.67)	2.72 (1.22)	2.38 (1.46)

Note: Numbers represent means (standard deviations).

Affidavit

Ich erkläre hiermit, dass ich die vorgelegten und nachfolgend aufgelisteten Aufsätze selbstständig und nur mit den Hilfen angefertigt habe, die im jeweiligen Aufsatz angegeben oder zusätzlich in der nachfolgenden Liste aufgeführt sind. In der Zusammenarbeit mit den angeführten Koautoren war ich mindestens anteilig beteiligt. Bei den von mir durchgeführten und in den Aufsätzen erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der Satzung der Justus-Liebig-Universität Giessen zur Sicherung guter wissenschaftlicher Praxis niedergelegt sind, eingehalten.

Giessen, 31.07.2023

____

Martin Haupt

Eingereichte Artikel:

- Wannow, S., Haupt, M. and Ohlwein, M. (2023), "Is brand activism an emotional affair? The role of moral emotions in consumer responses to brand activism", *Journal of Brand Management,* online first. https://doi.org/10.1057/s41262-023-00326-9

- Haupt, M., Wannow, S., Marquardt, L., Graubner, J.S. and Haas, A. (2023), "Who is more responsive to brand activism? The role of consumer-brand identification and political ideology in consumer responses to activist brand messages", *Journal of Product & Brand Management*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JPBM-10-2022-4193

- Haupt, M., Wannow, S., Freidank, J. and Haas, A., “AI wrote this? Effects of AI-powered content creation on employer attractiveness”, (working paper under review at *European Journal of Marketing*)

- Haupt, M., Freidank, J. and Haas, A., “Using Human-AI Collaborative Authorships to Escape Algorithm Aversion in Content Marketing”, (working paper in revision at *Review of Managerial Science*)
- Haupt, M., Rozumowski, A., Freidank, J. and Haas, A., “Seeking empathy or suggesting a solution? Effects of chatbot messages on service failure recovery to optimize digital platform efficiency” (working paper in revision at *Electronic Markets*)