Brands in Unsafe Places: Effects of Brand Safety Incidents on Brand Outcomes

Abstract

Well-publicized digital media incidents, in which brand content appears adjacent to "unsafe" content (e.g., negative content that is offensive, harmful, or uncomfortable), highlight the potential risk to a brand's reputation every time it advertises on digital platforms. Even as content moderation algorithms improve, brands cannot control digital environments fully, making it imperative for marketing managers to develop brand safety processes to keep a brand's reputation safe within *digital* advertising ecosystems, among their risk mitigation efforts. The current research accordingly attempts to establish when brand safety concerns are more or less likely to arise, according to specific consumer-, brand-, and incident-related moderators; why consumers react negatively to incidents, depending on their capacity to erode consumer trust in brands; and how and to what extent these combined elements affect various brand-related outcomes. Across data from Twitter (X) and six experiments, the authors distinguish brand safety incidents from other types of brand risks that demand managerial attention, and they empirically showcase how digital brand safety incidents influence consumers' attitudes and behaviors, as well as advertisers' outcomes. Building on these empirical findings, this article provides concrete, evidence-based suggestions for how to mitigate incidents, both before and after their occurrence.

Keywords: digital media, social media, brand safety, trust, digital advertising

In broad media environments, brands do not want to be associated with content that is inappropriate, offensive, controversial, or just even inconvenient in nature, such that it could potentially harm their reputation or performance. For example, cruise lines would prefer that their advertisements not appear next to articles about contagious airborne illnesses (Hsu 2020). Brands' attempts to limit such issues then determine their strategic *brand safety* initiatives. Although brand safety concerns are not new, contemporary digital environments pose novel and expanded challenges, as well as greater risk implications than arise in traditional media channels. Due to the rapid pace at which content, including brands' advertisements, gets served to audiences; the widespread reliance on algorithms to serve up digital content, without any human intervention to determine what content gets displayed to whom and when; and the sheer amount of available content, including user-generated content, brands simply cannot guarantee that their digital advertising always appears in entirely appropriate—that is, *safe*—digital media settings.

Increased brand safety risks in turn require increased efforts to keep the overall brand reputation safe in digital advertising ecosystems (Interactive Advertising Bureau [IAB] 2018; Johnson, Voorhees, and Khodakarami 2023). In such contexts, managing brand safety often entails *adjacency* considerations that stem from the perceived safety or suitability of content that a brand appears near, next, or adjacent to in a given media channel. Across digital environments, user-generated posts appear above or below brands' advertisements in Facebook and Instagram feeds; pre- and mid-roll advertisements run during videos on YouTube, which also features banner ad overlays. But the content that the brand appears adjacent to largely is beyond the brand's control, because it gets determined by opaque algorithms. In the struggle to ensure brands do not appear alongside inappropriate—that is, *unsafe*—content, companies devote substantial time, effort, and advertising dollars to try to secure safe ad space. In particular,

marketers have innovated various tools, webinars, and services that promise greater digital brand safety (Johnson, Voorhees, and Khodakarami 2023). Unfortunately though, such efforts have not proven widely successful, so marketers continue to seek (and offer to pay premium prices for) digital media environments that provide greater safety for their brand advertising. Not only is managing brand safety difficult in practice, with strategically important consequences, but it also involves substantial complexity, because the definition of "safe" adjacent content varies across brands, products, and audiences.

Some limited brand safety research has identified the effects of adjacent, unsafe digital content on critical brand outcomes such as ad recall, brand liking, and loyalty (Bushman 2007; Johnson, Voorhees, and Khodakarami 2023; Lee, Kim, and Lim 2021; Manatt, Avital, and Ofer 2018), though Bellman et al. (2018) suggest null effects. Notably though, prior research takes a generalized view of the overall negative versus positive environments surrounding brand advertisements. Such an approach fails to account for how various brand safety incidents, in different digital environments, involving diverse brands and consumers, might influence relevant brand outcomes. With this research, we seek to both establish that brand safety incidents occur and address these research gaps.

Using controlled experimental scenarios and data from real-world brand safety incidents, we explore a key mechanism that appears especially relevant to brand outcomes in risky settings: diminished perceptions of a brand's trustworthiness. To help managers design brand safety strategies, we also specify when safety incidents are more or less likely to affect brand outcomes, based on the incident type, brand features, and consumer beliefs. By investigating digital brand safety incidents from multiple perspectives (e.g., both consumers who experience them firsthand and those who hear about the incidents from other sources), we can better establish when and

why digital brand safety becomes a pressing issue. In addition, we build on and expand insights gathered from diverse literature pertaining to adjacency, proximity, contagion, and spillover effects. Leveraging these insights, we propose a novel process by which adjacent digital content affects downstream brand outcomes, through the erosion of brand trust. Our empirical findings reveal how digital brand safety incidents influence consumers' attitudes and behaviors, as well as the downstream consequences for advertisers, while also accounting for several potential moderators. We use a multimethod approach, in line with evidence that multimethod approaches are critical for examining brand safety (Johnson, Voorhees, and Khodakarami 2023). Finally, for managerial practice, we detail the unique effects for different digital media platforms, types of safety incidents, and brands of various sizes, reputations, and industries. Noting the wide prevalence of brand safety concerns, we also propose suggestions for how brands can mitigate them in digital media environments. These evidence-based suggestions highlight specific scenarios in which brand marketers should be more or less concerned about safety incidents.

Conceptual Framework

Brand Safety

Among varied perspectives on brand reputation management and brand risks, brand safety represents a distinct form of reputational risk that has not received a lot of attention in extant research. In Table 1 (and an expanded literature review in Web Appendix A), we provide explicit comparisons of brand safety with other types of brand risk, to highlight how it is similar to and distinct from other brand risks (e.g., product recalls, scandals, and brand spillover). In our proposed conceptual framework, brand safety represents a distinct phenomenon that requires specific consideration and that exerts unique effects on consumer behavior, relative to other types of brand risks.

By focusing particularly on brand safety, we can identify risks that demand increased efforts to keep the brand's overall reputation safe, particularly in digital advertising ecosystems (IAB 2018; Johnson, Voorhees, and Khodakarami 2023). Even as programmatic algorithms become more widely available and sophisticated, brand safety risks in digital advertising ecosystems remain external to the brand and outside its control, because the brand's advertisements appear adjacent to uncontrollable user-generated content. In this sense, brand safety concerns are clearly distinct from other reputational risks that are internal to a brand (e.g., product recalls) or over which they can exert greater potential control (e.g., customer complaints). Brand safety considerations also are distinctive in digital channels, and digital advertising in particular (cf. other channels or marketing domains), as is evident in extant industry and research definitions. Therefore, they cannot be addressed fully by research that investigates traditional advertising channels or spillover effects due to perceived adjacency in domains other than advertising.

Table 1: Distinguishing Brand Safety from Some Example Related Concepts

Study	Crisis Type	Advertising Focused?	Crisis Source	Brand Control	Digital
Bellman et al. (2018)	Negative content adjacency	Yes	External	Low	No
Borah and Tellis (2016)	Spillover and product recall	No	External	Low	No
Grégoire, Tripp, and Legoux (2009)	Customer complaints	No	Internal	High	Yes
Hansen, Kupfer, and Hennig-Thurau (2018)	Various crises	Yes	Internal	High	Yes
Knox and van Oest (2014)	Customer complaints	No	Internal	High	No
Koschate-Fisher, Hoyer, and Wolframm (2019)	Co-brand spillovers	No	External	Low	No
Lei, Dawar, and Lemmink (2008)	Product harm	No	Internal	Low	No
Patterson, Cowley, and Prasongsukarn (2006)	Service failure	No	Internal	High	No
Roehm and Tybout (2006)	Product harm	No	Internal	High	No
Smith, Bolton, and Wagner (1999)	Service failure	No	Internal	High	No
Srinivasan and Sarial-Abi (2021)	Algorithmic error	Yes	Internal	High	Yes
Votolato and Unnava (2006)	Spillover and crisis responses	No	External	Partial	No
The current study	**Brand safety**	**Yes**	**External**	**Low**	**Yes**

Our focus on brand safety also reflects the recognition that it represents an increasingly pressing concern for firms (Graham 2023), prompting many leading brands to establish dedicated brand safety offices, though little research focuses specifically on this critical concern.

Johnson, Voorhees, and Khodakarami (2023, p. 2) establish an initial definition of brand safety, as a "marketing strategy to ensure that online marketing investments are not served in environments (i.e., websites, videos, social media streams) that conflict with a brand's image." With a survey of marketing managers, they also report that 81.1% of the professionals agreed that brand safety was a major concern. In turn, these authors explicitly call for continued research to validate brand safety threats, determine how long related events have negative effects for firms, and identify potential moderating effects.

In our attempt to address these calls, we focus on digital domains, because in such contexts, brand safety is inherently connected to adjacency. The beneficial outcomes sought through brand advertisements easily might be undermined if those advertisements appear adjacent to unsafe or unsuitable content. We simultaneously (1) recognize that the mechanisms by which contamination, contagion, or spillover effects threaten brands in various types of media channels differ from but also (2) we leverage insights from prior research in those related domains to derive some initial, potential explanations of why unsafe adjacent content is likely to evoke brand-adverse effects. For example, physical proximity creates conditions for negative contagion and contamination effects in various consumer contexts, such that the desirability of an object decreases when it appears nearer in physical proximity to an undesirable contamination source (e.g., gore), even if the participants receive explicit information that no physical contact ever occurred (Kim and Kim 2011). We apply this notion to digital channels to predict that consumers might make associations across adjacent content (advertisements included), regardless of whether explicit associations exist.

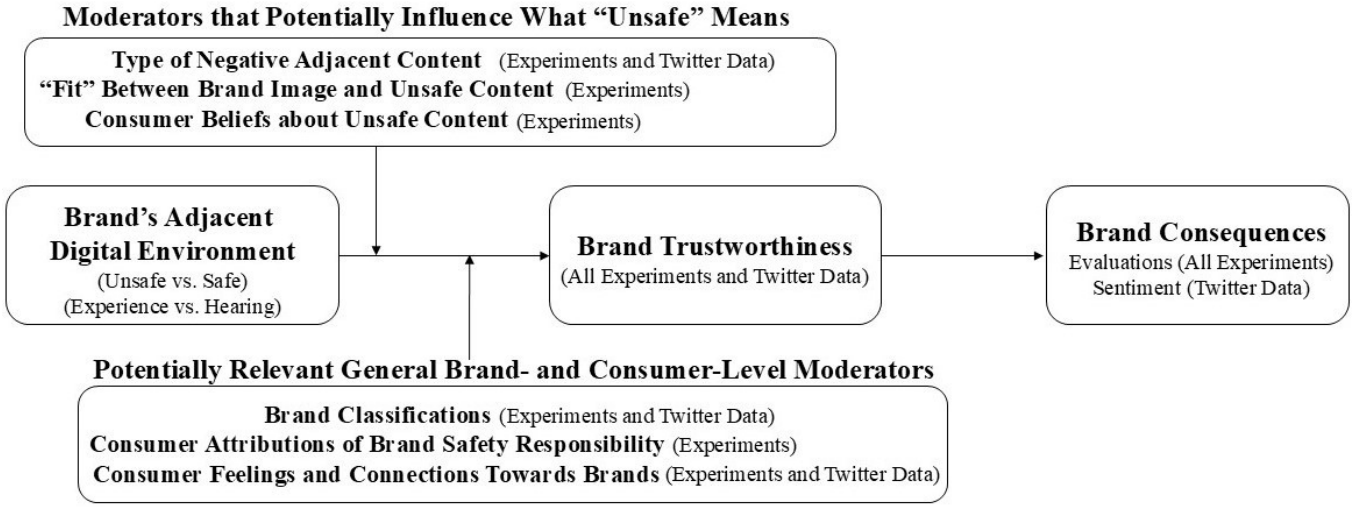
Furthermore, from brand advertising research conducted in non-digital channels, we leverage the notion that content adjacency and proximity affect relevant outcomes. For example,

in traditional television and print media, the tone and emotion of the content appearing before or after advertisements can affect consumers' brand perceptions (Nian, Hu, and Chen 2021). Unsafe content in television programs (e.g., violence, sex) can lower ad recall and brand intentions (Bushman 2005, 2007). In research into *digital* brand advertising, which is more directly relevant to our research, we also note evidence of positive effects of congruency between ads and content, such that consumers prefer to see matched content (Aribarg and Schwartz 2020; Belanche, Flavian, and Perez-Rueda 2017). When adjacent digital content is misaligned, brand unsafe, unsuitable (e.g., automobile ad adjacent to videos of a car crash), or generally unsafe (e.g., content about a school shooting), consumers' brand attitudes diminish, and some consumers even believe the ad placement was purposeful or represents an endorsement of the unsafe content (Manatt, Avital, and Ofer 2018). Pre-roll ads that appear before unsafe video content also can harm brand attributions (e.g., responsibility) and consumer intentions (e.g., recommendations; Lee, Kim, and Lim 2021).

Such adjacency and proximity effects highlight how brands (and related products) can be influenced by adjacent content, even if they are not formally linked. Building from these research streams (Web Appendix A), we introduce a conceptual framework to delineate what brand safety is, and what it is not. In turn, we use this foundation to derive the proposed empirical framework in Figure 1, which details how the safety of the digital environment in which brand advertising appears affects multiple brand outcomes, including consumer evaluations and brand sentiment. In detail, because consumers tend to perceive physically adjacent content as related, even if it comes from different sources, we posit that when digital advertisements appear adjacent to unsafe content, consumers' trust in the advertised brand decreases, which lowers those consumers' brand-related attitudes and associated behaviors. This erosion of trust also might be

moderated by various factors, related to the brand safety incident, brand classifications, and consumer-level perceptions and individual differences that influence trust in the brand. We delineate the other elements of this framework in the following sections.

Figure 1: Conceptual and Empirical Framework



The Erosion of Brand Trust

A brand's success depends powerfully on its perceived trustworthiness. Even when a brand suffers some negative associations, it can thrive if it previously has established its trustworthiness among consumers (Power, Whelan, and Davies 2008). Yet as industry-oriented research indicates, consumer trust has declined to low levels (Gallup 2023). In turn, and as supported by a recent meta-analysis (Khamitov et al. 2024), brand managers have critical roles to play to improve this critical metric. Noting the clear relevance of research that can identify the antecedents of trust perceptions, as well as marketing tactics for mitigating threats to brand trust, we investigate specifically how brand safety incidents, in conjunction with other drivers, influence brand trust, as an essential perception that in turn determines brand outcomes.

Trust in a source of information stems from perceptions of that source as reliable, consistent, and in possession of integrity (Doney and Cannon 1997; Morgan and Hunt 1994). These perceptions are based on inferences about the validity of expressed opinions. Brand trust specifically encompasses beliefs about the reliability and integrity of a brand, which indicate whether it can and will fulfill its commitments (Garbarino and Johnson 1999; McKnight, Choudhury, and Kacmar 2002). Such trust is particularly influential in uncertain settings (Doney and Cannon 1997; Moorman, Zaltman, and Deshpande 1992). Thus, we anticipate its centrality when consumers must decide whether to accept brand claims when they also encounter information that undermines their confidence (Bart et al. 2005).

Building on studies that describe spillover processes due to attribution and diagnosticity (Koschate-Fischer, Hoyer, and Wolframm 2019), we propose that consumers use adjacent unsafe content as information about the values embraced by a brand and its potential to fulfill promises (e.g., claims made). Therefore, if brand advertising appears adjacent to offensive content in an unsafe environment, consumers doubt its values, claims, or ability to follow through—all of which evoke reputational harm and greater uncertainty. Once they infer some connection between the unsafe content and brand advertising, audiences sense uncertainty about what the brand "stands for" and diminished trust. These feelings threaten downstream brand attitudes and behaviors (e.g., purchase intentions, loyalty, sales, word of mouth; Bart et al. 2005; Chaudhuri and Holbrook 2001). To understand these complex influences in the context of brand safety incidents, we consider several theoretically and managerially supported moderators and boundary conditions that likely interact with unsafe digital environments to influence audiences' perceptions of and trust in the brand.

Defining Unsafe Content

Types of negative adjacent content. Defining what constitutes inappropriate adjacent content, and thus what kinds of content threaten brand safety, arguably becomes particularly nuanced and complicated in digital environments. Even if we accept the standard "Dirty Dozen" typology (IAB 2018), not all unsafe content is equal. Therefore, we turn to negative publicity research (Cleeren, van Heerde, and Dekimpe 2013), which identifies crisis severity as an important moderator (Johar, Birk, and Einwiller 2010), and propose that reactions to incidents differ with the type of unsafe content. That is, consumers' reactions generally are more negative when the crises are severe (Liu and Shankar 2015). Therefore, we anticipate that different brand safety incidents (e.g., hate speech, explicit content, depictions of tobacco or marijuana usage) exert different effects.

Some unsafe content probably is considered unsafe by most or all consumers (e.g., content promoting terrorism). But using a programmatic approach to address all unsafe content is problematic, because it ignores brand and consumer traits that make some unsafe content more or less suitable, relative to the brand. Digital environments feature vast amounts of potentially negative content, not all of which is unsafe for brands. Consider content that depicts tobacco usage. According to the Dirty Dozen (IAB 2018), it is always unsafe, so programmatic brand safety strategies would enforce rules that require avoidance of such content. A brand advertisement that appears next to a scientific post about the harmful consequences of smoking is unlikely to suffer the same negative reactions than if it appears next to a young user's post depicting and glorifying their smoking habit though. In the first scenario, consumers are less likely to question the integrity of the brand than in the second, so brand trust and outcomes should be less affected by this form of negative adjacent content, which is not really "unsafe."

Therefore, we predict that the type of negative content that is adjacent to a brand's content might play a moderating role on brand safety outcomes.

Consistency with the brand's image and type. Similarly, not all content, even that which reflects the Dirty Dozen categories, imposes equivalent effects. Rather, the impacts likely depend on the fit between a brand's image and the negative adjacent content. As research has shown, positive congruency between ads and adjacent content evokes improved consumer perceptions (Aribarg and Schwartz 2020; Belanche, Flavian, and Perez-Rueda 2017). In turn, for some brands, specific types of unsafe content could be regarded as a good fit. A brand that has partnered strategically with celebrities and influencers known to smoke probably be unharmed by appearing adjacent to even unsafe content related to marijuana use; it even might benefit from the perceived fit with this content. We will empirically examine this perceived fit or congruency between "unsafe" content and brands to determine if consistency is a moderating factor for consumer perceptions after brand safety incidents.

Consumers' beliefs about unsafe content. From consumers' perspective, the personal relevance of, or beliefs, surrounding an unsafe topic should influence how they process adjacent digital content (Huh and Reid 2007; Lee, Kim, and Lim 2021; Whelan and Dawar 2016). For example, self-relevant information is chronically more accessible (Naylor, Lamberton, and Norton 2011), so someone who encounters unsafe content that seems relevant to them may express stronger reactions, leading to more severe negative effects. Perhaps female consumers sense greater alienation when confronted with sexist or misogynistic content next to an advertisement for cosmetics. Alternatively, self-relevance could mitigate brand safety effects; a smoker might perceive content related to vaping as a form of personalized advertising (which consumers generally view positively; Lambrecht and Tucker 2013), rather than a marker of

unsafe content. Thus, we predict that consumer-level factors also might moderate the perceived trustworthiness and downstream outcomes of brands after being exposed to unsafe content.

Brand- and Consumer-Related Characteristics

Brand classifications. The type of brand offering affects how consumers integrate external information into their decision-making. Experiential (vs. material) goods, such as services, tend to be more subjective, which makes them harder to compare (Holbrook and Hirschman 1982). Thus, service-oriented, experiential brands might be more subject to consumers' biases or heuristic cues, whereas for product-focused brands, consumers can base their decisions on available, objective information. In turn, consumers might value objective product reviews more than they do subjective service reviews (Dai, Chan, and Mogilner 2020). When consumers can access objective information, heuristic cues, such as those evoked by unsafe adjacent content surrounding brand advertising, should be less impactful. Therefore, both product- and service-oriented brands need to address brand safety incidents, but service brands may be at risk of more severe negative outcomes.

Another distinction refers to utilitarian versus hedonic brands (Chen, Lee, and Yap 2017; Kronrod and Danziger 2013). Consumers generally indicate weaker connections to utilitarian brands and are more likely to switch, whereas hedonic brands tend to engender more positive emotions and stronger bonds (Keller 2001; Keller and Lehmann 2006). This categorization also influences how consumers engage in information searches; they allocate more time to social media when considering hedonic purchases (Chung et al. 2023; Li et al. 2020). Such a tendency should increase the likelihood that they are exposed to brand safety incidents. We predict that perceptions of a brand as utilitarian or hedonic thus should interact with the brand safety incident to influence downstream outcomes, though the direction of this effect is unclear *a priori*.

Consumers' brand attributions, connections, and feelings. Consumer-level differences are relevant for predicting adjacency and spillover effects; we predict they also likely influence perceptions of brand safety incidents. For example, consumers can attribute blame for a safety incident to the brand, the platform, or algorithms. Spillover research indicates that attributions of responsibility affect both whether a spillover occurs and its implications for perceived brand equity (Koschate-Fischer, Hoyer, and Wolframm 2019). Perceptions of attribution have also been examined as a downstream implication of brand safety (Lee, Kim, and Lim 2021; Manatt, Avital, and Ofer 2018). As a reflection of how consumers feel about a brand, self–brand connections also determine whether negative brand incidents function like threats to the consumer's sense of self (Escalas and Bettman 2003). Consumers with stronger such connections express more anger in response to a brand crisis (Mosley, Schweidel, and Zhang 2024). Yet consumers' commitment to and relationships with a brand also might minimize the threats created by negative brand information for important outcomes like brand equity (Ahluwalia, Unnava, and Burnkrant 2001).

Finally, factors such as consumer familiarity, awareness, prior experience, and liking influence brand perceptions related to trust, equity, and downstream outcomes (Garbarino and Johnson 1999, Keller 1993). Therefore, we predict that these preexisting perceptions also influence how consumers process brand safety information. The direction of these effects is not clear though. Different types of consumers attend to different types of information, which might lead to confirmation biases or overweighting of specific information. Such outcomes can cause non-obvious effects, reflecting each consumer's prior perceptions of the brand in the context of a brand safety incident (Dawar and Pillutla 2000). If consumers take the information conveyed by a brand safety incident at face value, without considering positive balancing or competing

information, the incident might exert a stronger influence on their attitudes and behaviors than otherwise. Such impacts also may be more intense for brands that consumers like less or have less familiarity with, or they could be more powerful for brands for which consumers have positive associations. For example, positively viewed brands might be judged more harshly for an apparent integrity lapse and face worse backlash. As it is unclear if these factors will moderate, and in what direction they may impact brands, we will examine these various factors as possible moderators of brand safety outcomes across experiments.

Overview of Studies

Using both data gathered from Twitter (now X) related to real brand safety incidents and controlled experiments, we examine our core prediction that brand safety incidents can negatively impact brands, and that it occurs through a process related to the lowered perceived trust in the brand. Additionally, using measured and manipulated moderators, we examine the potential impacts of potential moderators as detailed in the framework in Figure 1.

First, utilizing Twitter data from real brands that have suffered public brand safety incidents, we showcase initial evidence of brand safety's real negative impact in the field. We also examine our process and moderators. With the experiments, we then confirm those initial findings in preregistered and controlled scenarios wherein we show that brand safety incidents exert negative brand effects. As well, the results of the mediation and moderation analyses provide evidence of the predicted trust-related process. Lastly, we examine alternative process explanations, extensions, and possible boundaries of these effects (see Table 2 for a summary of experimental findings). This combined empirical approach allows us to highlight theoretically and managerially relevant moderators that can inform ongoing brand safety and risk assessment

strategies. The experimental data, syntax, PDFs of the experiments, and preregistrations are publicly available at https://osf.io/rsxet/?view_only=eff67e8ce7cd45fbbb5eb678bbce718e.

Table 2: Summary of Key Experimental Results

Study	Sample Size + Population	Manipulated Moderator	Measured Variables for Potential Moderation	Key Conditions Influencing Brand Evaluations		Main Finding(s) on Brand Trust (M) and Evaluations (DV) Detailed in the Manuscript and Web Appendix
				Unsafe Environment M (SD)	Safe Environment M (SD)	
2a	296 Prolific Academic		Consumer Demographics Smoking Status Online Behaviors	3.78 (1.22)	4.25 (1.21)	When a brand's ad was adjacent to unsafe content (vs. a control of no adjacent content), trust in the brand and brand evaluations were both lower.
2b	299 MTurk with CloudResearch		Consumer Demographics Brand: Familiarity, Liking, Usage, Self-Brand Connection, Commitment	3.90 (1.37)	4.80 (1.20)	When a brand's ad was adjacent to unsafe (vs. safe) content, trust in the brand and brand evaluations were both lower. Brand liking was a significant moderator of our effect (WA E).
2c	379 Prolific Academic		Consumer Demographics Brands: Familiarity, Liking, Usage, Self-Brand Connection, Commitment	Experience: 3.70 (1.42) Hearing: 3.73 (1.46)	4.53 (1.36)	When a real brand's ad was adjacent to unsafe (vs. safe) content, trust in the brand and brand evaluations were both lower for participants who saw the unsafe content first hand, and for those who read about the incident. Political orientation, brand liking (as in study 2b), and brand commitment were found to be marginally, or fully significant moderators (WA E).
3	398 Prolific Academic		Consumer Demographics Self-Brand Connection, Attributions of Responsibility	3.42 (1.34)	4.15 (1.22)	Provides evidence that trust is the strongest underlying process for our effect compared to a number of theoretically and managerially other constructs tested.
WAN	431 Prolific Academic	Baseline Brand Trust	Consumer Demographics Liking and Familiarity of Products and Brands Self-Brand Connection	Low Trust: 2.99 (1.53) High Trust: 3.62 (1.49)	Low Trust: 4.42 (1.34) High Trust: 4.49 (1.38)	Using real brands that differ on consumer trust, and showcasing real brand ads over videos, we moderate our effect wherein a highly trusted brand has attenuated effects to a brand safety incident (compared to a lesser trusted brand).

4a	327 Prolific Academic		Consumer Demographics Brands: Familiarity, Liking, Usage, Self-Brand Connection, Commitment	Unsafe: 3.80 (1.58)	Negative Safe: 4.47 (1.19) Control Safe: 4.49 (1.33)	When a real brand's ad was adjacent to unsafe (vs. negative but safe and purely safe) content, trust in the brand and brand evaluations was lower. This study differentiates the negative brand outcomes from brand safety incidents from emotional contagion. Gender, brand commitment for Nissan, and brand commitment for Disney were found to be marginally, or fully significant moderators (WA E).
4b	537 CloudResearch Panels	Brand Fit with Unsafe Content	Consumer Demographics Smoking Beliefs	Low Fit: 4.25 (1.41) High Fit: 4.21 (1.64)	Low Fit: 4.69 (1.27) High Fit: 4.26 (1.37)	Our effect is moderated by the perceived fit between unsafe content and the brand whose ad is adjacent to it. Both brand trust and brand evaluation effects are attenuated in the high fit brand conditions (vs. in the low fit conditions). Age and acceptability beliefs regarding cannabis were significant moderators (WA E).
WA O	559 CloudResearch	Timing of Brand Response	Consumer Demographics Attributions of Responsibility Smoking Beliefs	***All Unsafe*** *No PR Response:* 4.04 (1.45) *Immediate Response:* 4.77 (1.35) *Soon After Response:* 4.59 (1.44) *Weeks After Response:* 4.33 (1.44)		Brands that wait to respond after brand safety incidents are judged more harshly. However, even a late response can be seen as better than no response at all—highlighting that brands should respond, and do so as soon as possible, once an incident becomes known. Significant moderation was found for brand's perceived responsibility for events, and a consumers' preexisting beliefs about cigarettes and marijuana. Brands needed to respond to safety incidents in a timely (immediate or within a day) manner but there were the highest chances of backfiring effects when brands wait too long to respond to incidents.
WA P	588 Prolific Academic	Secondary Brand Ad Response	Consumer Demographics Attributions of Responsibility Self-Brand Connection	***All Unsafe at Time One*** *Control:* 3.48 (1.28) *High Liking:* 3.79 (1.33) *High Awareness:* 3.80 (1.26) *High Credibility:* 3.75 (1.27)		After a brand safety incident, a subsequent brand advertisement can mitigate the negative effects on brand evaluations by emphasizing some positive brand perceptions that were seen in the Twitter data and some experiments to protect brands (e.g., likability, awareness, credibility). Both platform and algorithm responsibility interacted with how positively certain messaging was perceived after a brand safety incident and a consumer's age was also a significant moderator.

Web Appendix B includes all stimuli and pretest or manipulation check analyses across experiments. Web Appendix C lists each item participants saw, in the order shown, for all experiments. Web Appendix D showcases a table for all experiments which includes all potential measured moderators as covariates. None of our results significantly change with their inclusion

or exclusion. In Web Appendix E we report the findings when each potential measured moderator is independently tested as a moderator of the effect of brand safety (and other manipulated moderators) on brand evaluations across experiments. When significance is found, we highlight these results in the relevant main study discussions and in the General Discussion.

Study 1: Initial Evidence from Twitter

This study pursues two main objectives: to establish exploratory and descriptive evidence of the impact of real-world brand safety incidents on how consumers think about and discuss brands in digital media environments, and to provide preliminary evidence of some of the moderators identified in our conceptual model (Figure 1). The evidence we gather from the Twitter data is descriptive; with the controlled experiments in Studies 2–4, we provide rigorous tests of causality. For our empirical context, we focus on Twitter (now X). At the time of our data collection, Twitter provided a suitable source of public data for assessing how consumers' brand attitudes change over time; tweets often revealed users' brand perceptions (Culotta and Cutler 2016) and could be used to measure a brand's reputation over time (Rust et al. 2021). We know of no database that keeps track of brand safety incidents, so with the assistance of a global marketing trade association that, at the time, focused on helping marketing leaders tackle brand safety risks for their brands, we gathered a set of incidents and corresponding brands. For each brand in our analysis, we required multiple sources (e.g., news articles, popular press) that confirmed the brand safety incident, which increases the likelihood that consumers were aware of the incidents, even if they did not experience them personally. Then, for each brand that experienced such a brand safety incident, we determine the sentiment (positivity or negativity) of tweets that mention it, and track any changes in brand sentiment over short time windows, before and after the incident. We postulate that these changes in sentiment can be indicative of the

effect of safety incidents. In addition, we explore the role of trust, by tracking if brand trust decreased after the incident.

Data

The data involve 86 brands affected by safety incidents, as identified by the global marketing trade association. In Web Appendix F, Table WF1, we provide detailed information, including examples of media coverage surrounding each incident and relevant dates. Some dates vary slightly for different brands involved in the same incident. For each brand, we identify the unique date that the first few tweets about the incident appeared and affirm that the time lags between the initial media coverage and the appearance of discussions on Twitter were reasonable. For each brand incident, we define the time period after the date that tweets first appear as an "after brand safety incident" (ABSI) period and the days prior to this date as the "before" period, when brand-related conversations should be unaffected by the incident. With fixed windows of the same length for both time periods, we test if brand sentiment changed after compared with before the incident.

Specifically, we used the Python package SNSCRAPE to collect tweets in these before and after periods and gathered all tweets that refer to the affected brand by name, seven days before and after the incident. The resulting data set contains 6,725,225 tweets, approximately half of which (52.01%) appear after the brand safety incident. In Web Appendix G, we list all brands, the volume of tweets, and mean sentiments in 7-day windows before and after the incident. A considerable cross-brand variance appears in the amount of attention brands receive, both in general and in the ABSI period.

Main effect on sentiment

We compute tweet sentiment with the Python package VADER, applied to cleaned tweets.[1] Among the tools available to calculate sentiment, VADER is explicitly designed for social media text and has been shown to outperform both individual human raters and benchmark machine learning models that calculate sentiment, especially on microblogging sites such as Twitter (Hutto and Gilbert 2014). It assigns an overall compound sentiment value, ranging from –1 (very negative) to +1 (very positive), to each tweet. We provide model-free evidence of the effect in Web Appendix G and anecdotal evidence of the effect in Web Appendix H.

We estimate the following fixed-effects regression model, in which individual tweet i pertaining to brand-incident j posted at time t is the unit of analysis. Each tweet's compound sentiment score $Sentiment_{ijt}$ serves as the dependent variable; the independent variable of interest is a dummy labeled $ABSI_{ijt}$ that indicates if the tweet appeared after the incident (1) or before (0). We control for potential relationships of tweet sentiment with characteristics of the Twitter user (numbers of followers and friends, $Followers_{ijt}$ and $Friends_{ijt}$) and the tweet (numbers of favorite ratings and retweets, $Favorites_{ijt}$ and $Rewteets_{ijt}$; whether the tweet is a reply or original, $Reply_{ijt}$; and if it contained a URL, URL_{ijt}). These controls are included in our model to account for systematic patterns across tweet or user characteristics and sentiment. Finally, we include a vector of fixed effects for the brand incident with γ_j, as well as a vector of time-specific fixed effects τ_t for the hour of the day, weekend, and month. The time windows range from 1 to 7 days. Thus,

$$Sentiment_{ijt} = \alpha + \beta_1 ABSI_{ijt} + \beta_2 Followers_{ijt} + \beta_3 Friends_{ijt} + \beta_4 Favorites_{ijt} + \beta_5 Retweets_{ijt} + \beta_6 Reply_{ijt} + \beta_7 URL_{ijt} + \gamma_j +$$

Table 3: Main Effects of Brand Safety Incidents on Brand Sentiment across Time Windows

DV: Compound	One-day	Two-day	Three-day	Four-day	Five-day	Six-day	Seven-day

[1] In the cleaning process, we removed terms that do not contribute meaningfully to the computation of sentiment, such as URLs, hyperlinks, and usernames.

sentiment of a tweet	Estimate	SE		Estimate	SE		Estimate	SE		Estimate	SE		Estimate	SE		Estimate	SE		Estimate	SE	
IVs																					
After brand safety incident (ABSI)	-0.0031	0.0008	***	-0.0063	0.0006	***	-0.0031	0.0005	***	-0.0027	0.0004	***	-0.0103	0.0004	***	-0.0108	0.0003	***	-0.0095	0.0003	***
Number of followers *(10^{-6})*	0.0158	0.0013	***	0.0135	0.0009	***	0.0148	0.0007	***	0.0162	0.0007	***	0.0082	0.0004	***	0.0063	0.0004	***	0.0046	0.0003	***
Number of friends *(10^{-6})*	0.3410	0.0284	***	0.4040	0.0206	***	0.4550	0.0176	***	0.4730	0.0155	***	0.5160	0.0142	***	0.5220	0.0133	***	0.5250	0.0123	***
Number of favorites *(10^{-6})*	-4.4700	5.3700		-1.5200	1.6300		-2.1100	1.9900		-2.3200	1.7600		-2.6800	2.0000		-2.3000	1.8900		-2.3200	1.7800	
Number of retweets *(10^{-6})*	14.2000	12.5000		8.2400	5.9100		3.7600	6.2800		6.7700	5.7800		8.2100	6.0500		6.9200	5.8600		6.6300	5.3800	
Tweet is a reply	0.0407	0.0015	***	0.0324	0.0011	***	0.0321	0.0009	***	0.0327	0.0008	***	0.0278	0.0007	***	0.0259	0.0006	***	0.0258	0.0006	***
Tweet contains URL	0.0437	0.0012	***	0.0470	0.0009	***	0.0458	0.0007	***	0.0471	0.0006	***	0.0534	0.0006	***	0.0537	0.0005	***	0.0523	0.0005	***
Brand incident FE	Yes			Yes			Yes			Yes			Yes			Yes			Yes		
Hour of day FE	Yes			Yes			Yes			Yes			Yes			Yes			Yes		
Weekend FE	Yes			Yes			Yes			Yes			Yes			Yes			Yes		
Month FE	Yes			Yes			Yes			Yes			Yes			Yes			Yes		
N	1,009,024			2,004,000			2,949,392			3,939,001			4,792,068			5,616,147			6,472,533		
R^2	0.0901			0.0896			0.0879			0.0891			0.0810			0.0799			0.0782		

$^{***}p < .01$, $^{**}p < .05$, $^{*}p < .1$.
Notes: DV = dependent variable, IV = independent variables, FE = fixed effect. The estimation is based on robust standard errors (SE).

The results in Table 3 indicate a statistically significant decrease in sentiment after the brand safety incident in all models. While the negative effect on sentiment appears to reach its greatest strength during the 5–6-day window, considering the descriptive nature of this study, we focus here on the significance and directionality of the effect rather than its magnitude.

Moderating effects on sentiment

To provide additional descriptive evidence in support of our conceptual framework, we explore how salient brand classifications, the type of negative adjacent content, and consumer brand feelings might moderate the demonstrated effects on sentiment. After considering the full set of moderators in our conceptual framework (Figure 1), we chose this subset in this study since they could be analyzed using the Twitter data. In our subsequent experiments, we corroborate these findings where possible and dive deeper into other moderators as well.

To ensure a close alignment of our descriptive results with the results from the experimental Studies 2–4, we focus our moderator analyses on 3-day windows on either side of

each brand safety incident. That is, in our experimental studies, we measure the effects of brand safety incidents on consumers immediately after the incidents, so using 3-day windows better captures the immediate aftermath of the incidents while also remaining robust to variations across brands, incidents, and time periods since considerable variability appears in the news coverage of various incidents and in the likelihood of some events being picked up in Twitter chatter after some delay. For robustness, we also report the parallel results using 2- and 4-day windows in Web Appendixes I–M.

We start by addressing the moderating role of three types of salient brand classifications. First, we consider the brand's predominant offering, whether it is product- or service-oriented. We assigned brands to these two categories using our own expert judgment and brand knowledge (Table WF1). To estimate the moderation in the regression model, we begin with the specification in Equation 1, then add the interaction between ABSI and a dummy variable that takes a value of 1 if a brand's predominant offering is service-oriented and 0 if it is product-oriented.[2] The results (see Column I of Table 4) affirm that service-oriented brands suffered significantly more negative sentiment after an incident than did product-oriented brands.

Second, we classify brands according to their utilitarian versus hedonic positioning. For this assessment, we surveyed participants from MTurk (N = 787, M_{age} = 32.7 years, 27.0% women) to gauge their perceptions. For each brand, multiple participants rated, on a five-point scale, the extent to which they perceived the brand to be predominantly hedonic (=1) or predominantly utilitarian (=5) (Web Appendix J). We obtained average scores for each brand and divided the brands in the data set, according to whether they scored above or below the average across all brands. To estimate this moderation, we added an interaction between ABSI and a

[2] Note that we do not include the main effect of this dummy variable, because it is simultaneous with the brand incident fixed effect.

dummy variable that takes a value of 1 if a brand's average utilitarian score is above the average score and 0 otherwise. The results (Column II of Table 4) affirm that brands with above-average utilitarian (hedonic) scores experienced a lesser (greater) impact of brand safety incidents. In other words, hedonic brands suffered more severe sentiment decreases after an incident than did utilitarian brands.

Third, as another brand classification moderator, we measure attitudinal brand perceptions, specifically brand familiarity and brand liking. Participants from MTurk (N = 388, M_{age} = 36.0 years, 31.9% women) rated, for each brand, their familiarity with and liking of the brand (Web Appendix K). We obtained average scores for each brand, then classified the brands according to whether they scored above or below the average across all brands. We estimate the moderating effect using the same method that we applied for brand positioning. The results (Columns III and IV, Table 4) show that brands that possess above (below) average familiarity and liking scores were affected less (more) by brand safety incidents.

Table 4: Moderators of the Effects of Brand Safety Incidents on Brand Sentiment

	(I)			(II)			(III)			(IV)			(V)			(VI)		
DV: Compound sentiment of a tweet	Brand Offering			Brand Type			Brand Familiarity			Brand Liking			Incident Type			Personal Connection		
	Estimate	SE		Estimate	SE		Estimate	SE		Estimate	SE		Estimate	SE		Estimate	SE	
IVs																		
ABSI	0.0068	0.0007	***	-0.0079	0.0007	***	-0.0088	0.0008	***	-0.0116	0.0008	***	-0.0001	0.0008		-0.0070	0.0007	***
ABSI × Service	-0.0176	0.0009	***															
ABSI × Utilitarian				0.0081	0.0009	***												
ABSI × Familiarity							0.0084	0.0010	***									
ABSI × Liking										0.0129	0.0009	***						
ABSI × Incident type: hate speech													-0.0041	0.0010	***			
ABSI × Personal connection																0.0073	0.0009	***
Number of followers *(10^{-6})*	0.0146	0.0007	***	0.0147	0.0007	***	0.0147	0.0007	***	0.0147	0.0007	***	0.0148	0.0007	***	0.0148	0.0007	***
Number of friends *(10^{-6})*	0.4550	0.0176	***	0.4550	0.0176	***	0.4550	0.0176	***	0.4550	0.0176	***	0.4550	0.0176	***	0.4550	0.0176	***
Number of favorites *(10^{-6})*	-2.0800	2.0000		-2.0900	1.9900		-2.0900	2.0000		-2.0700	1.9900		-2.1000	1.9900		-2.1000	1.9900	
Number of retweets *(10^{-6})*	3.7700	6.3100		3.7600	6.2900		3.7800	6.3000		3.7000	6.2800		3.7400	6.2900		3.7400	6.2800	
Tweet is a reply	-0.0321	-0.0009	***	0.0321	0.0009	***	0.0321	0.0009	***	0.0322	0.0009	***	0.0321	0.0009	***	0.0321	0.0009	***
Tweet contains URL	-0.0459	-0.0007	***	0.0457	0.0007	***	0.0458	0.0007	***	0.0458	0.0007	***	0.0458	0.0007	***	0.0458	0.0007	***

Brand incident FE	Yes	Yes	Yes	Yes	Yes	Yes
Hour of day FE	Yes	Yes	Yes	Yes	Yes	Yes
Weekend FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	2,949,392	2,949,392	2,949,392	2,949,392	2,949,392	2,949,392
R^2	0.0880	0.0879	0.0879	0.0879	0.0879	0.0879

***$p < .01$, **$p < .05$, *$p < .1$.
Notes: DV = dependent variable, IV = independent variables, ABSI = after brand safety incident, FE = fixed effect. The estimation is based on robust standard errors (SE).

Turning to the type of negative adjacent content in the brand safety incident, we conceptualize this moderator according to a categorization of the safety incidents. Our data feature multiple brand safety incidents, which we can assign to the Dirty Dozen framework (IAB 2018) of content categories that tend to be unsafe (e.g., violence, hate speech, adult content). We coded the incidents in the data set according to the descriptions of the nature of the incident published in news articles (Web Appendix F). We coded the incidents into two general types: (1) hate speech/acts of aggression and (2) adult or sexually explicit content. To test if content type moderates the effect of an incident, we include the interaction between ABSI and a dummy variable that takes a value of 1 for hate speech–related incidents and 0 for adult or sexually explicit content–related incidents (Web Appendix L). The coefficient of this interaction term is negative and significant (Column V, Table 4), such that sentiment was significantly more negative when incidents involved hate speech rather than adult content.

Finally, we include consumers' perceived self–brand connection. For each brand, we asked participants from MTurk (N = 787, M_{age} = 32.7 years, 27.0% women) to complete two items (1–5 scale) that refer to the extent to which they feel connected with the brand (Web Appendix M). We obtained average scores for each brand across the two items and divided the brands according to whether they ranked above (or below) the average score across all brands. To estimate the moderation, we add the interaction between ABSI and a dummy variable that takes a value of 1 if a brand's connection score is above the average and 0 otherwise. The results

(Column VI, Table 4) show that brands that evoked above (below) average connections were affected less (more) by brand safety incidents.

Trust

In our conceptual framework, the central process through which brand safety incidents affect brand outcomes is through an erosion of trust. To test the validity of this process with real-world data, we check if the likelihood that a tweet's emotional content is associated with brand trust weakens after a safety incident. For this analysis, we employ the Python package NADE (Natural Affect Detection; Hotz-Behofsits, Wlömert, and Abou Nabout 2025), which can infer basic emotions from social media messages. Specifically, NADE takes text as input (e.g., single tweet) and provides the intensity that this text is associated with eight basic emotions (Hotz-Behofsits, Wlömert, and Abou Nabout 2025). For the purposes of this study, we focus on the intensity that a tweet is associated with trust as a basic emotion; for each tweet, this intensity serves as the dependent variable. The regressions feature the independent variables and model specification from Equation 1. The results, as detailed in Table 5, consistently indicate that the intensity that a tweet is associated with trust drops significantly after a safety incident. While this drop appears to recover and become statistically indistinguishable from zero just a week after the incident, we stress that these findings are solely descriptive. Overall, following a brand safety incident, there appears to be a drop in trust expressed in tweets that mention the affected brand.

Table 5: Effect of Brand Safety Incidents on Trust across Time Windows

DV: Compound sentiment of a tweet	One-day		Two-day		Three-day		Four-day		Five-day		Six-day		Seven-day	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
IVs														
ABSI	-0.0004	0.0002 *	-0.0006	0.0002 ***	-0.0004	0.0001 ***	-0.0006	0.0001 ***	-0.0006	0.0001 ***	-0.0004	0.0001 ***	0.0001	0.0001
Number of followers *(10^{-6})*	0.0103	0.0006 ***	0.0088	0.0004 ***	0.0089	0.0003 ***	0.0096	0.0003 ***	0.0097	0.0003 ***	0.0087	0.0002 ***	0.0074	0.0002 ***
Number of friends *(10^{-6})*	0.1200	0.0091 ***	0.1240	0.0065 ***	0.1300	0.0055 ***	0.1250	0.0049 ***	0.1240	0.0045 ***	0.1280	0.0042 ***	0.1350	0.0039 ***
Number of favorites *(10^{-6})*	-3.0400	1.4500	-1.0400	0.9710	-1.4100	1.0300	-0.3280	0.6400	-13100	1.1800	-0.9020	0.9860	-0.9620	0.8710
Number of retweets *(10^{-6})*	1.7300	3.7100	-4.0700	2.7900	-3.2800	2.5900	-2.5400	1.7700	-1.7900	2.8500	-2.2400	2.3900	-1.5500	1.8200

Tweet is a reply	0.0244	0.0004 ***	0.0207	0.0003 ***	0.0202	0.0002 ***	0.0210	0.0002 ***	0.0248	0.0002 ***	0.0240	0.0002 ***	0.0231	0.0002 ***
Tweet contains URL	0.0804	0.0003 ***	0.0819	0.0002 ***	0.0820	0.0002 ***	0.0823	0.0002 ***	0.0857	0.0002 ***	0.0860	0.0001 ***	0.0858	0.0001 ***
Brand incident FE	Yes		Yes		Yes		Yes		Yes		Yes		Yes	
Hour of day FE	Yes		Yes		Yes		Yes		Yes		Yes		Yes	
Weekend FE	Yes		Yes		Yes		Yes		Yes		Yes		Yes	
Month FE	Yes		Yes		Yes		Yes		Yes		Yes		Yes	
Year FE	Yes		Yes		Yes		Yes		Yes		Yes		Yes	
N	1,009,024		2,004,000		2,949,392		3,939,001		4,792,068		5,616,147		6,472,533	
R^2	0.1652		0.1714		0.1713		0.1707		0.1758		0.1755		0.1746	

***$p < .01$, **$p < .05$, *$p < .1$.
Notes: DV = dependent variable, IV = independent variables, FE = fixed effect. The estimation is based on robust standard errors (SE).

Moderating effects on trust

Similar to our brand sentiment analyses, we explore if and how salient brand classifications, type of negative adjacent content, and consumer brand feelings might moderate the effects on trust. The results of these analyses, which replace brand sentiment with trust as the dependent variable (Table 6), remain directionally consistent. We report the results using 2- and 4-day windows in Web Appendixes I–M.

Table 6: Moderators of the effect of Brand Safety Incidents on Trust

DV: Compound sentiment of a tweet	Brand Offering		Brand Type		Brand Familiarity		Brand Liking		Incident Type		Personal Connection	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
IVs												
ABSI	0.0008	0.0002 ***	-0.0049	0.0002***	-0.0040	0.0002***	-0.0026	0.0002***	0.0005	0.0002*	-0.0016	0.0002 ***
ABSI X Service	-0.0022	0.0003 ***										
ABSI X Utilitarian			0.0075	0.0003***								
ABSI X Familiarity					0.0054	0.0003***						
ABSI X Liking							0.0034	0.0003***				
ABSI X Incident type: hate speech									-0.0012	0.0003***		
ABSI X Personal connection											0.0022	0.0003 ***
Number of followers *(10^{-6})*	0.0089	0.0003 ***	0.0088	0.0003***	0.0088	0.0003***	0.0088	0.0003***	0.0089	0.0003***	0.0089	0.0003 ***
Number of friends *(10^{-6})*	0.1290	0.0053 ***	0.1290	0.0055***	0.1280	0.0055***	0.1290	0.0055***	0.1280	0.0055***	0.1290	0.0055 ***
Number of favorites *(10^{-6})*	-1.4000	1.0300	-1.3900	1.0300	-1.4000	1.0300	-1.4000	1.0300	-1.4000	1.0300	-1.4000	1.0300
Number of retweets *(10^{-6})*	-3.2700	2.5900	-3.2700	2.5900	-3.2700	2.5900	-3.2900	2.5900	-3.2800	2.5900	-3.2800	2.5900
Tweet is a reply	0.0202	0.0002 ***	0.0202	0.0002***	0.0202	0.0002***	0.0202	0.0002***	0.0202	0.0002***	0.0202	0.0002 ***
Tweet contains a URL	0.0820	0.0002 ***	0.0819	0.0002***	0.0819	0.0002***	0.0820	0.0002***	0.0820	0.0002***	0.0820	0.0002 ***
Brand-incident FE	Yes		Yes		Yes		Yes		Yes		Yes	
Hour of day FE	Yes		Yes		Yes		Yes		Yes		Yes	
Weekend FE	Yes		Yes		Yes		Yes		Yes		Yes	
Month FE	Yes		Yes		Yes		Yes		Yes		Yes	
N	2,949,392		2,949,392		2,949,392		2,949,392		2,949,392		2,949,392	
R^2	0.1713		0.1715		0.1714		0.1713		0.1713		0.1713	

****p* < .01, ***p* < .05, **p* < .1.
Notes: Robust standard errors were used in the estimation.

Discussion

Study 1 provides preliminary, real-world support for the predicted effect; it also offers some initial, descriptive evidence of the pertinent roles of some of the moderators proposed in our conceptual model. Yet we acknowledge that these findings might be alternatively explained, due to unobserved variables (e.g., source of initial information, whether consumers experienced the incident themselves). Future research could focus on using observational data to rigorously establish causality and quantify the effect size. Our investigation of moderating effects, to the extent possible, suggests brands that evoke greater familiarity, liking, and personal connection experience minimal negative impacts of brand safety incidents, implying the risks of brand safety incidents may be more conditional than absolute. For brands with weaker reputations, safety incidents may erode trust and amplify negative effects, whereas stronger brands may be

somewhat “protected” from such outcomes. We call for continued research into exploring the validity of such mechanisms with observational data. In the following studies we build upon this initial evidence through a series of randomized, controlled, preregistered experiments.

Study 2: Main Effect Findings Related to Brand Safety Incidents

In Study 2, building on the findings in Study 1, we seek to establish experimentally that brand safety incidents can negatively impact multiple brand dimensions (e.g., purchase intentions, brand liking). We integrate these dimensions into a brand evaluation index. Across three preregistered experiments, conducted on different social media platforms (Facebook, X, and YouTube), involving different types of brands (hypothetical or real; product- or service-oriented), and reflecting different consumer experiences (actual experience vs. hearing about incidents), we investigate multiple brand safety incidents that represent distinct IAB (2018) classifications. We consistently find that being associated with unsafe content can harm brand evaluations, due to the negative effects on perceived trustworthiness.

Study 2a: Unsafe Versus No Adjacent Content (Product Context)

Participants, method, and design. Participants were recruited on Prolific Academic (preregistered: https://aspredicted.org/DHY_8FT) and invited to participate in an experiment with a two-cell (safety: unsafe, control) between-subjects design (N = 296, after 4 people were excluded for failing the attention check; 42.6% women, 55.7% men, 1.7% other or prefer not to state, M_{age} = 36.40 years). They completed a Facebook task, following the explanation that they would see a screenshot from Facebook, then answer some questions about it. In the unsafe condition, the screenshot featured a Facebook newsfeed that included stimuli related to the effects of vaping on health, above an advertisement for a fictional clothing company called Anderson & Co. In the control condition, no adjacent content appeared near the brand ad on the

newsfeed. Next, participants rated Anderson & Co. on five items in an overall brand evaluation scale (α = .93; e.g., "I would consider buying products from the brand"; 1 = "strongly disagree," 7 = "strongly agree"). They also completed a brand trust item ("I believe the brand is trustworthy"), in randomized order within the same block as the other items. Finally, participants answered an attention check item and provided basic demographic information (age, gender, Facebook usage, time spent online, and smoking status).

Brand evaluations and trust. We regressed brand evaluations on brand safety (unsafe = -1, control = 1). The overall model ($F(1, 294) = 11.12$, $p = .001$) and the effect of brand safety (b = .24, $t(294) = 3.34$, $p = .001$) were significant. Participants in the control condition reported higher brand evaluations (M = 4.25, SD = 1.21) than those in the unsafe condition (M = 3.78, SD = 1.22), and the brand trust results were consistent (overall: $F(1, 294) = 6.03$, $p = .015$; safety: b = .17, t(294) = 2.46, $p = .015$; $M_{control} = 4.49$, SD = 1.16; $M_{unsafe} = 4.14$, SD = 1.25).

Mediation. In the PROCESS macro (model 4, 10,000 resamples; Hayes 2017), we tested for mediation, using brand trust as the mediator, brand safety as the independent variable, and brand evaluation as the dependent variable. Higher safety (i.e., control condition) positively predicted brand trust (b = .17, t = 2.46, $p = .0146$), and brand trust positively predicted brand evaluations (b = .70, t = 16.68, $p < .001$). The confidence interval (CI) for the indirect effect of trust did not include 0 (b = .12, SE = .05, CI_{95}[.02, .22]), indicating the presence of mediation.

Study 2b: Unsafe Versus Safe Adjacent Content (Services Context)

Participants, method, and design. Participants recruited on MTurk with CloudResearch features (preregistered: https://aspredicted.org/K9J_G7D) were invited to participate in an experiment with a two-cell (safety: unsafe, safe) between-subjects design (N = 299, after 2 people were excluded for failing the attention check; 53.2% women, 45.8% men, 1% other or

prefer not to state, M_{age} = 41.13 years). Participants learned that they would see a screenshot from Twitter (X), then answer some questions about it. In the unsafe condition, the screenshot featured a Twitter feed that included stimuli related to recent anti-Semitic behavior in the United States, above a real recent advertisement for AT&T, the telecommunications company. In the safe condition, the adjacent content was a post about upcoming holidays at the time of data collection (specifically, Hanukkah, to keep the religious component consistent across conditions). After seeing the randomly assigned content, participants rated AT&T on the same evaluation scale as in Study 2a (α = .95) and the same trust item. They also answered an attention check item, indicated their feelings toward AT&T (familiarity, experience using, liking, commitment to, connection with), and provided basic demographic information.

Brand evaluations and trust. We regressed brand evaluations on brand safety (unsafe = -1, safe = 1). The overall model ($F(1, 297) = 35.97$, $p < .001$) and the effect of brand safety (b = .45, t(297) = 6.00, $p < .001$) were both significant. Participants who saw the brand's ad adjacent to safe (vs. unsafe) content reported higher brand evaluations (M_{safe} = 4.80, SD = 1.20; M_{unsafe} = 3.90, SD = 1.37). A similar pattern emerged for brand trust (overall: $F(1, 297) = 27.74$, $p < .001$; safety: b = .44, t(297) = 5.27, $p < .001$; M_{safe} = 4.90, SD = 1.39; M_{unsafe} = 4.03, SD = 1.47).

Mediation. We used the PROCESS macro (model 4, 10,000 resamples; Hayes 2017) to test for mediation, with brand trust as the mediator, brand safety as the independent variable, and brand evaluation as the dependent variable. Higher safety positively predicted brand trust (b = .44, t = 5.27, $p < .001$), and brand trust positively predicted brand evaluations (b = .79, t = 30.75, $p < .001$). The CI for the indirect effect of trust did not include 0 (b = .34, SE = .07, CI_{95}[.21, .48]), indicating the presence of mediation.

Study 2c: Generalization of the Effect (Experiencing vs. Hearing)

We propose that consumers might not need to experience unsafe content firsthand for negative brand outcomes to arise. The Study 1 findings suggest that brands suffer diminished sentiment after a brand safety incident, which presumably reflects online chatter more broadly, not just comments from consumers who experienced the incident. Accordingly, we predict that hearing about unsafe content in connection with a brand can cause negative effects that are consistent with actually experiencing a brand safety incident, such that third-party reports might create seemingly adjacent associations.

Participants, method, and design. For this study, we recruited 450 participants on Prolific Academic (preregistered: https://aspredicted.org/KTG_C92) and invited them to participate in an experiment with a three-cell (safety: unsafe-experience, unsafe-hearing, control) between-subjects design (N = 351 after preregistered exclusions; 53.8% men, 44.7% women, 1.5% other or did not want to disclose; M_{age} = 36.19 years). Participants were randomly assigned to complete a task in which they learned about a real brand safety incident on YouTube. In the unsafe-experience condition, participants saw a screenshot of a real co-branded ad from Nissan and Disney that promoted the Nissan Leaf and the release of the film *A Wrinkle in Time*, played in the mid-roll spot for an actual video of former KKK grand wizard (and self-avowed Nazi) David Duke. That video, when released in 2018, prompted significant news attention to brand safety issues on digital platforms. In the unsafe-hearing condition, participants read a description of Duke's video and the brand safety incident but did not see an image of the ad playing over the video itself. In the control condition, people saw only the screenshot of the brand ad, without any information about the video on which it appeared on YouTube.

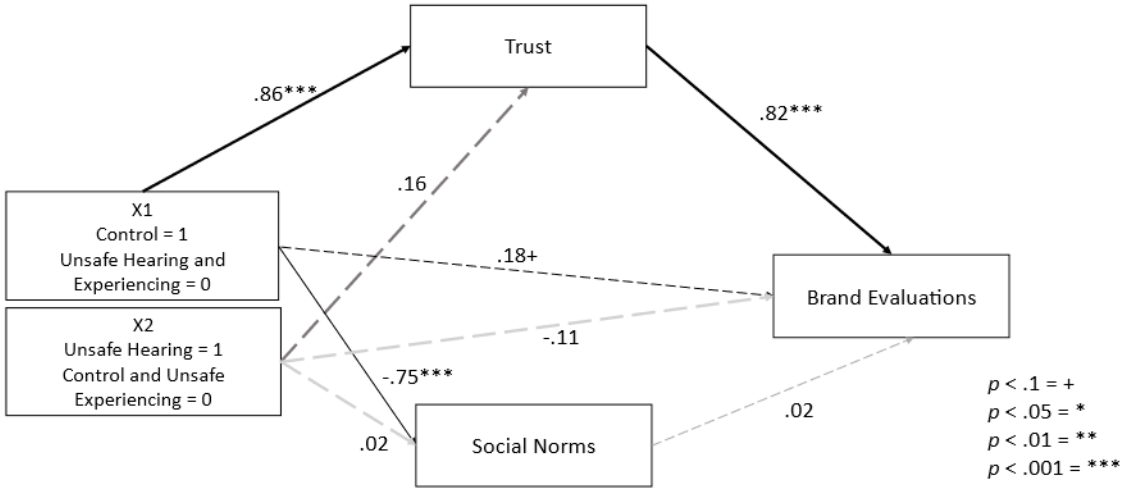
After seeing the randomly assigned content, participants rated Nissan in an overall brand evaluation scale, as in Studies 2a and 2b (α = .96), and on the same trust item. In addition, they offered responses to five items pertaining to perceived social norms for online content (α = .75). Participants completed an attention check item, provided demographic information (age, gender, time spent online, political orientation), and indicated what their feelings toward Nissan and Disney had been, prior to the survey (familiarity, usage, liking, commitment to, connection to).

Brand evaluations and trust. We ran an analysis of variance (ANOVA) for brand evaluations. The overall model was significant ($F(2, 348) = 13.67$, $p < .001$), such that participants who saw the brand advertisement without any adjacent content reported higher brand evaluations (M = 4.58, SD = 1.37) than those in the unsafe-experience (M = 3.70, SD = 1.42; $p < .001$) and unsafe-hearing (M = 3.73, SD = 1.46; $p < .001$) conditions. The two unsafe conditions did not differ significantly ($p = .891$). The ANOVA for trust similarly revealed that the overall model was significant ($F(2, 348) = 10.07$, $p < .001$). Participants who saw the brand advertisement without adjacent content reported higher trust (M = 4.58, SD = 1.47) than those in the unsafe-experience (M = 3.72, SD = 1.48; $p < .001$) and unsafe-hearing (M = 3.88, SD = 1.56; $p < .001$) conditions, and again, the two unsafe conditions did not significantly differ ($p = .413$).

Mediation. Using the PROCESS macro (model 4, 10,000 resamples; Hayes 2017) to test for mediation, we include brand trust as the mediator, brand safety as the independent variable, and brand evaluation as the dependent variable. For the multicategorical mediation analysis, we rely on indicator coding: X1 represents the control condition compared with the two unsafe conditions, and X2 refers to the unsafe-hearing condition compared with the other two conditions (see Figure 2). The results affirm that X1 positively predicted brand trust (b = .86, t = 4.17, $p < .001$), and brand trust positively predicted brand evaluations (b = .82, t = 33.20, $p < .001$). The

CI for the indirect effect of trust for X1 did not include 0 (b = .71, SE = .17, CI_{95}[.38, 1.05]). In contrast, all the X2 effects were non-significant, and its CI spanned 0 (b = .13, SE = .16, CI_{95}[-.19, .45]). In a competing mediation model, with perceived social norms and brand trust as parallel mediators, we do not find any significant indirect effects of X1 or X2 for the social norm items, and the indirect effect through trust remained significant for X1, as predicted (b = .71, SE = .17, CI_{95}[.38, 1.04]).

Figure 2: Multicategorical Mediation through Trust and Social Norms



Discussion

Studies 2a–2c offer evidence consistent with the findings of Study 1: Brand safety incidents adversely affect brands, whether the comparison involves a lack of adjacent content or safe content. The adverse effects arise among consumers who experience the incidents firsthand and those who only read about a brand being adjacent to unsafe content. If emotion or affect informed the primary process, we would expect the outcomes to differ between consumers who experience the unsafe content firsthand or just learn about it from some dispassionate, external outlet. Many consumers learn about brand safety incidents from third parties, and in this sense, we predict that the process that drives brand evaluations is more cognitively (versus affectively)

driven. Additionally, the lack of significant indirect effects from perceived social norms (Study 2c) enables us to rule it out as an alternative explanation. Across these studies, we also establish that brand safety effects appear at least partly due to an erosion of consumers' trust in the brand.

We additionally begin to examine some potential moderators that are based in consumer beliefs regarding the unsafe content or the brand impacted by the incident (see Figure 1). Across Study 2b and 2c (i.e., studies with real brands), we find that brand liking and brand commitment stand out the most as potential moderators. In Study 2c we also see moderation due to political orientation. These consumer demographic finding may imply that certain consumers view "unsafe" content as more or less "unsafe" due to their beliefs generally or specifically about brands. We will continue to examine these types of consumer feelings towards brands and unsafe content through our experiments and elaborate on all findings and patterns in Web Appendix E.

Study 3: Alternative Explanations

Although Study 2 provides evidence that brand trust underlies the effects of brand safety on evaluations we also recognize consumer phenomena tend to reflect multiple determinants (Pham 2013). Thus, with Study 3, we examine some other potential processes that we have not yet tested, as suggested by prior literature. First, studies of affect and contagion suggest that generalized emotive states, such as moods, can transfer to and influence subsequent judgments (Pham 1998), and the emotion evoked by persuasive communications can influence nearby, unrelated content, in line with affect-as-information theory (Hasford, Hardesty, and Kidwell 2015). Second, we measure attributions of blame and responsibility to the brand, platform, and underlying algorithm, then determine if such perceptions of responsibility matter, reflecting findings that suggest attributions of responsibility determine spillover effects (Koschate-Fischer, Hoyer, and Wolframm 2019; Lee, Kim, and Lim 2021; Manatt, Avital, and Ofer 2018). Third,

information consistency (with surrounding content or within an ad) can influence ad effectiveness and consumers' attention to the information (Sahni and Nair 2020). The perceived diagnosticity of event information and a consumer's involvement with or attention to messages also might shift, positively or negatively (Ahluwalia, Unnava, and Burnkrant 2001). We note mixed findings regarding the impact of controversial content (which might be considered unsafe) on brand attitudes and memory of brand information: Bellman et al. (2018) find no effects, but Bushman (2007) indicates that unsafe content hinders memory. Fourth, brand commitment might constrain negative spillover effects, so we measure self–brand connection as an individual difference moderator (Escalas and Bettman 2003) and perceived brand commitment as a potential process mechanism (Ahluwalia, Unnava, and Burnkrant 2001).

Participants, Method, and Design

We recruited 399 participants on Prolific Academic (preregistered: https://aspredicted.org/DDP_PBD) for a single-factor, two-cell (brand safety: unsafe, safe), between-subjects design (N = 397; 2 people were excluded for failing the attention check; 57.2% men, 40.3% women, 2.6% other or prefer not to answer; M_{age} = 39 years). They completed a YouTube task. In the unsafe condition, the screenshot represented a video about vaping at children's concerts, with a banner ad for the fictional clothing company (Anderson & Co.) overlaid at the bottom of the screen. In the safe condition, the screenshot reflected a video about a coffee shop, and the same ad was overlaid in the same position. Participants indicated their brand evaluation s (α = .94) and their brand trust, as in prior experiments.

They next completed items, in random order, that tested potential alternative explanations: responsibility for ad placement (three separate items measuring responsibility attributions to brands, websites, or algorithms), affect as information (measured two ways: how

the ad makes participants feel on five dimensions such as happy and sad [α = .96] and if participants perceive the information as for themselves or others [α = .93]), mood (taking three negative mood items and subtracting three positive mood items from the negative mood; α = .89), ad involvement (three bipolar scale items; α = .92), information consistency between the ad and adjacent content (two bipolar scale items; α = .91), advertisement information diagnosticity (three bipolar scale items; α = .90), and brand commitment (three items about loyalty and brand switching willingness; α = .47[3]). We also collected self–brand connection (α = .95) as a general individual difference moderator. At the end, participants completed the attention check and provided demographic information.

Results and Discussion

Brand evaluations. We regressed brand evaluations on brand safety (unsafe = -1, safe = 1). The overall model was significant ($F(1, 395) = 32.88$, $p < .001$), as was the effect of brand safety (b = .37, t(395) = 5.734, $p < .001$), such that being adjacent to safe (vs. unsafe) content predicted higher evaluations ($M_{safe} = 4.15$, SD = 1.22; $M_{unsafe} = 3.42$, SD = 1.34).

Mediation. Brand safety significantly affected all potential mediators (all $p < .025$) except attributions of responsibility, which were insignificant (all $p > .30$). Thus, we tested the items predicted by brand safety simultaneously, to establish a conservative test (Buechel and Janiszewski 2014).[4] Using the PROCESS macro (model 4, 10,000 resamples; Hayes 2017), in a regression with brand safety and all mediators as predictors, we find that brand trust positively predicted evaluations (b = .65, $t = 17.86$, $p < .001$), as did information diagnosticity (b = .14, $t = 3.89$, $p = .001$) and brand commitment (b = .10, $t = 4.39$, $p < .001$). No other tested mediators

[3] Despite the low alpha of this scale, we include it as it was a preregistered alternative explanation, using only two items vs. three does not produce a stronger alpha, and it is a theoretically relevant measure to consider.

[4] Multicollinearity tests confirm that including all the predictors in the model simultaneously does not cause variance inflation issues (all variance inflation factors < 2.04).

predicted brand evaluations (all $p > .17$). All three significant pathways offered evidence of mediation too (trust: b = .23, SE = .04, CI_{95}[.15, .32], diagnosticity: b = .05, SE = .02, CI_{95}[.02, .08], brand commitment: b = .02, SE = .01, CI_{95}[.003, .04]; Table 7).

Table 7: Mediation Results, Study 3

	Effect of Brand Safety on M				Y (Brand Evaluations)				
	Coeff.	SE	*t*	*P*	Coeff.	SE	*t*	*p*	Indirect Effects (CI)
X (Safety)	---	---	---	---	.0220	.0401	.5478	.5841	---
M_1 (Brand Trust)	.3561	.0624	5.7039	.0000	.6541	.0366	17.8615	.0000	.1489,.3222
M_2 (Others Brand Trust)	.3542	.0712	.0000	.0000	.0077	.0288	.2683	.7886	-.0165,.0245
M_3 (Incidental Affect)	-.4397	.0561	7.8340	.0000	-.0387	.0391	-.9892	.3232	-.0119, .0494
M_4 (Ad Involvement)	.1995	.0870	2.2938	.0223	.0140	.0236	.5937	.5531	-.0069, .0144
M_5 (Information Consistency)	.4134	.0794	5.2077	.0000	.0398	.0290	1.3745	.1701	-.0117, .0455
M_6 (Info Diagnosticity)	.3380	.0730	4.6278	.0000	.1360	.0350	3.8905	.0001	.0166, .0833
M_7 (Mood)	-.5054	.1167	-4.3298	.0000	-.0189	.0179	-1.0526	.2932	-.0094, .0323
M_8 (Brand Commitment)	.1894	.0843	2.2482	.0251	.1044	.0238	4.3882	.0000	.0027, .0422
Constant	---	---	---	---	.1134	.2530	.4481	.6543	---
					R^2 = .7096				
					F(9, 387) = 105.08, $p < .0001$				

Discussion

Even if unsafe (vs. safe) brand environments significantly affect consumer emotions, attitudes, and beliefs, brand trust still emerges as the strongest explanatory measure and a key factor in our findings in Study 3. Notably, we find weaker mediation through information diagnosticity and brand commitment. Considering that diagnosticity has been described as a component of trust (Mizerski 1982), it seems logical that this related explanation is significant, even if it exerts a weaker effect than trust.[5] We did not predict mediation through brand commitment, with the reasoning that both commitment and trust are central to relationship marketing (Morgan and Hunt 1994), but it stands to reason that as trust increases, commitment may increase in tandem (even if to a lesser extent).

Study 4: Boundary Conditions of the Adverse Effects of Negative Unsafe Adjacent Content

[5] As an additional test of the process, we ran a process by moderation study (https://aspredicted.org/5WX_BDS), with a 2 (brand safety: unsafe, safe) × 2 (brand trust: low, high) between-subjects design (after preregistered exclusions, N = 431; 54.3% men, 45.7% women; M_{age} = 35 years). We manipulated the baseline level of trust in the advertised brand (i.e., highly trusted versus untrusted brand, according to industry reports). When controlling for other explanatory factors, the results indicate that a consumer's trust in a brand influenced the degree to which the brand safety incident proved harmful (for full details, see Web Appendix N).

With Studies 4a and 4b, we test for potential boundary conditions. Prior research predicts the chances of spillover or contamination from negative content to other unrelated stimuli (Hasford, Hardesty, and Kidwell 2015; Kim and Kim 2011), and in Study 4a, we examine if brand safety concerns are unique to negative content that is deemed specifically unsafe or unsuitable versus that which is just generally negative. In Study 4b, we elaborate on these findings by examining how the fit among a brand, its advertising, and its target consumers might influence the strength of the effects of brand safety incidents on consumer outcomes.

Study 4a: Incident Type—Unsafe versus Safe but Negative

Participants, method, and design. With 441 participants recruited through Prolific Academic for this preregistered study (https://aspredicted.org/BTZ_JCX), we conducted a three-cell between-subjects study (brand safety: unsafe negative, safe negative, control). After the preregistered exclusions, 327 participants remained (45.9% men, 49.5% women, 4.6% other or prefer not to say; M_{age} = 37.46 years). They completed a task similar to the one in Study 2c. In the unsafe condition, participants saw the screenshot of a real co-branded ad from Nissan and Disney, played in the mid-roll spot of the video of David Duke. In the safe-negative condition, participants saw the same ad, but it played over a somber YouTube video that discussed medical racism and statistics regarding maternal death rates based on race. In the control condition, no context was provided about where the ad appeared on YouTube.[6]

Participants rated Nissan on five items in an overall brand evaluation scale, as in Study 2c (α = .96), and evaluated brand trust. Then they indicated their sense of how unsafe and inappropriate the content was, over which the ad played (11 items; α = .99) and how that content made them feel (5 items from Study 3; α = .92). Finally, participants answered attention check

[6] The unsafe condition seemed significantly more unsafe than the negative but safe and control conditions (both $p < .001$). Both unsafe and negative conditions prompted significantly lower affect than the control (both $p < .002$).

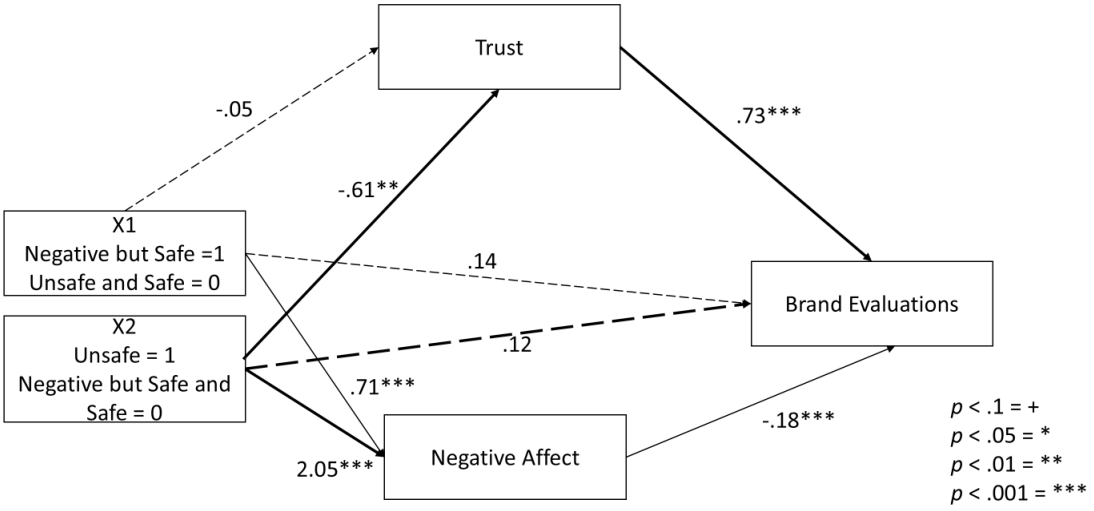
items and provided basic demographic information, along with their feelings toward Nissan and Disney related to commitment, liking, and familiarity (as in Study 2c).

Brand evaluations and trust. According to an ANOVA, the overall model was significant ($F(2, 327) = 8.41$, $p < .001$); participants who saw the brand's ad adjacent to the unsafe content reported lower brand evaluations (M = 3.80, SD = 1.58) than those in the safe but negative condition (M = 4.47, SD = 1.19; $p = .001$) and the control condition (M = 4.49, SD = 1.33; $p < .001$). The safe and negative condition did not significantly differ from the control condition ($p = .912$). The ANOVA for trust consistently revealed the significance of the overall model ($F(2, 327) = 5.43$, $p = .005$). Participants who saw the brand's ad adjacent to the unsafe content reported lower trust (M = 4.00, SD = 1.79) than those in the safe but negative condition (M = 4.56, SD = 1.32; $p = .009$) and the control condition (M = 4.61, SD = 1.34; $p = .002$). The negative and control conditions did not significantly differ ($p = .798$).

Mediation. We used the PROCESS macro (model 4, 10,000 resamples; Hayes 2017) to test for mediation, with brand trust as the mediator, brand safety as the independent variable, and brand evaluation as the dependent variable. In the multicategorical mediation analysis, X1 represented the safe but negative condition, compared with the other two conditions, and X2 represented the unsafe condition, compared with the other two conditions (see Figure 3). In this study, X2 negatively predicted brand trust (b = -.61, t = -3.09, $p = .002$), and brand trust positively predicted brand evaluations (b = .82, t = 33.80, $p < .001$). The CI for the indirect effect of trust for X2 did not include 0 (b = -.50, SE = .18, CI_{95}[-.84, -.15]), indicating the presence of mediation. All the X1 effects were non-significant, and the CI spanned 0 (b = -.04, SE = .15, CI_{95}[-.33, .24]). In a competing mediation model, with negative affect as a parallel mediator, the indirect effects through negative affect for both X1 (b = -.13, SE = .04, CI_{95}[-.21, -.06]) and X2

(b = -.36, SE = .09, CI_{95}[-.56, -.19]) were significant, whereas the indirect effect through trust was significant for only X2 (b = -.45, SE = .16, CI_{95}[-.76, -.14]).

Figure 3: Multicategorical Mediation through Trust and Negative Affect



Discussion. Replicating prior findings, we find that participants in the unsafe condition expressed poorer brand evaluations than those in the safe conditions. Yet this effect of negative unsafe content arises only if the brand ad appears adjacent to truly unsafe (vs. just affectively negative) content. Appearing proximate to negative content that is not necessarily unsafe does not appear to harm brand evaluations, despite its influence on consumers' evaluations. While this may seem counter to prior research (Hasford, Kidwell, and Hardesty 2015), we propose that this may be in part due to the contextual difference in this research (e.g., generalized priming tasks vs. digital environments). Future research can continue delving into the types of negative content that cause emotional contagion effects versus more cognitive outcomes.

Additionally, this study identified some moderating effects of consumer characteristics (i.e., gender), and brand connection and commitment. As detailed in Web Appendix E, our findings somewhat overlap, but are not identical, to all of the moderation tests run in Study 2c which uses the same safe and unsafe conditions. Interestingly, in line with some research that

shows that those with the highest brand connections may express more anger in response to a brand crisis (Mosley, Schweidel, and Zhang 2024), our moderation findings in both Studies 2c and 4a exhibit an instance of a backfiring effect or spillover effect wherein commitment or connection to a brand partner (Disney) significantly impacted evaluations to the focal brand (Nissan). Future research can continue to unpack how all possible moderating variables relating to consumers, incidents, and brands can interact with one another (versus in isolation) to explain when brand safety incidents will be most severe.

Study 4b: Congruency of Unsafe Content—Is All Unsafe Content Equal?

Participants, method, and design. For this preregistered study (https://aspredicted.org/98X_VRH), we recruited 613 participants from CloudResearch panels. It featured a 2 (brand safety: unsafe, control) × 2 (brand fit: low, high) between-subjects design (after preregistered exclusions,[7] N = 537; 38% men, 61.5% women, .6% other or prefer not to say; M_{age} = 51.61 years). Participants completed a YouTube task. In the unsafe conditions, they saw the screenshot of a video discussing smoking marijuana and cannabis; in the safe conditions, the neutral screenshot came from a video that played ambient white noise. In all four conditions, an ad overlay appeared for a fictional fast-food restaurant, SnackShack, with images of the kinds of food it sold and a tagline. In the high fit conditions, the tagline mentioned, "Midnight Munchies: Satisfying Cravings, One Bite at a Time!" and the brand was discussed as being for young adults, fulfilling late night cravings, and being associated with recreational smokers. In the low fit conditions, the tagline promised, "Family-Style Flavor: Every Bit, a Moment of

[7] We excluded participants if they failed two attention checks, in line with the preregistration. If we removed all participants who failed only one check, the sample would shrink to 327 people, and cell sizes would be less than 100. We cannot explain why the attrition was higher for this study than for the other studies; it might be due to the generally older demographic profile of participants on this online platform. The core findings of the focal pairwise contrasts remain statistically similar, using either sample. We prefer to analyze the larger sample here, to ensure sufficient power for the interactions and moderated mediation analyses.

Togetherness!" and the brand was described as targeting families, bringing people together, and being associated with parents wanting to encourage more family togetherness.

After seeing the randomly assigned content, participants rated the brand on the evaluation scale ($\alpha = .96$) and trust item from our prior studies. They completed the attention checks and provided basic demographic information (age, gender, religious beliefs, thoughts on the appropriateness of smoking cigarettes and weed), along with a measure of how unsafe they perceived the content to be (6-item index, $\alpha = .95$).

Brand evaluations and trust. We ran an ANOVA of brand evaluations, including brand safety (unsafe = -1, safe = 1), brand fit (low = -1, high = 1), and their interaction. The main effect of brand safety was significant ($F(1, 533) = 3.92$, $p = .048$), brand fit was marginally significant ($F(1, 533) = 3.63$, $p = .057$), and their interaction was directionally consistent ($F(1, 533) = 2.65$, $p = .104$). According to preregistered planned contrasts, when participants saw an ad for a low fit brand, the results replicated our prior findings: Brand evaluations were worse when an ad overlaid an unsafe (vs. safe) video ($M_{unsafe} = 4.25$, SD = 1.41; $M_{safe} = 4.69$, SD = 1.27; $p = .012$). However, for the high fit brand there were no significant differences due to video safety ($M_{unsafe} = 4.21$, SD = 1.64; $M_{safe} = 4.26$, SD = 1.37; $p = .802$).

In the ANOVA for trust, brand safety was marginally significant ($F(1, 533) = 3.35$, $p = .068$), brand fit was not significant ($F(1, 533) = 1.61$, $p = .204$), and their interaction was significant ($F(1, 533) = 4.21$, $p = .041$). The preregistered planned contrasts indicated that when participants saw the low brand fit ad, their trust diminished if they saw an ad over an unsafe (vs. safe) video ($M_{unsafe} = 4.29$, SD = 1.35; $M_{safe} = 4.77$, SD = 1.29; $p = .007$). We again found no significant differences in trust for the high fit brand, regardless of video safety ($M_{unsafe} = 4.38$, SD = 1.65; $M_{safe} = 4.36$, SD = 1.41; $p = .875$).

Moderated mediation. In the moderated mediation analysis (PROCESS Model 8, 10,000 resamples; Hayes 2017), brand safety was the independent variable, brand fit was the moderator, brand trust was the mediator, and brand evaluations was the dependent variable. It revealed a significant index of moderated mediation (b = -.21, SE = .10, CI_{95} [-.41, -.012]). The conditional indirect effect of brand safety on evaluations, through trust, was positive and significant in the low fit condition (b = .20, SE = .07, CI_{95} [.07, .33]), and it was not significant in the high fit condition (b = -.01, SE = .08, CI_{95} [-.16, .14]).

Discussion. Replicating our prior findings, participants in the unsafe condition expressed poorer brand evaluations than those in the safe condition, but only if the brand itself evoked low (vs. high) fit perceptions. In this sense, Study 4b reveals a managerially relevant boundary condition: congruency or perceived fit between the brand and the unsafe content. Additionally, we found that consumer perceptions towards cannabis acceptability significantly moderated our results. Similar to the political orientation finding in Study 2c and the gender moderation found in Study 4a, this finding may imply that beliefs that consumers hold generally can influence how unsafe content is seen as being; moderating the strength of brand safety effects (see Web Appendix E). These results also reemphasize the Study 4a findings, namely, that not all negative content is equal, and all unsafe content should not be treated the same way by brands. The type of unsafe content and perceived fit should both be used to inform brands' online advertising strategies and potential brand safety mitigation tactics.

General Discussion

In practice, pulling brand advertising completely from digital environments is not a viable way to deal with brand safety concerns. Relying on generalized programmatic or algorithmic systems also appears ineffective, as demonstrated by high-profile brand safety

incidents. Instead, brands need detailed strategies to manage and mitigate the risks surrounding brand safety incidents, which should account for the influences of various factors on both consumer- and market-level reactions. Our experiments and real-world data help shed some new light on how brand trust can mediate—and various consumer, brand, and incident factors can moderate—the extent to which appearing adjacent to unsafe content affects brands. In presenting the theoretical and practical implications of our findings, we also indicate how the contributions derived from the current research suggest avenues for continued investigations.

Theoretical Contributions

Some prior experimental research has predicted that unsafe content affects ad recall, attitudes, and intentions (Lee, Kim, and Lim 2021; Manatt, Avital, and Ofer 2018). With a multimethod approach, we provide more comprehensive tests of brand safety incidents, involving a wide variety of unsafe content scenarios, consumers who experience unsafe content or just learn about the incidents secondhand, and a broad array of managerially relevant outcomes. Brands do not necessarily suffer adverse consequences among every consumer, every time their advertising happens to appear adjacent to some unsafe content in digital environments (Studies 4a and 4b; see also Bellman et al. 2018). The variance in outcomes suggests that other factors, in tandem with safety incidents, determine the extent of harm, which implies substantial research potential and options for building on our findings.

Our consideration of the process by which the identified effects emerge is distinctive (brand trust versus other processes, such as affect or responsibility), as is our inclusion of theoretically and managerially relevant moderators suggested by prior research. Thus far, brand safety research has focused on establishing main effects, not *why* brands might be negatively affected by unsafe content (Bellman et al. 2018; Manatt, Avital, and Ofer 2018). However,

digital environments invariably create unique risk for different brands and consumer audiences, reflecting suitability concerns and individual consumer beliefs about safe content. Therefore, mitigation efforts should center on building (proactively) or reestablishing (reactively) trust. Even if it is not the only relevant factor (Study 3), trust exerts a consistently strong influence on the downstream outcomes of safety incidents for different consumers and scenarios, and it is something that brands can work to improve. To support such efforts, continued research might identify when alternative processes may be more or less impactful than brand trustworthiness, especially when used to define proactive and reactive mitigation efforts.

The framework of moderators we propose is theoretically novel and expands on research from multiple, disparate, but related literature streams. Because the framework accounts for numerous potentially relevant brand-, incident-, and consumer-related moderators, our research extends prior studies that explore one or two such influences at a time. In turn, we theoretically and empirically depict how these components can mitigate negative brand safety effects. We hope continued research will adopt and apply this insightful approach as well.

Managerial Implications

The results that we derive across our multimethod studies have notable implications for managers. Broadly, the moderators we test can be classified as factors over which brand managers have little to no control (immutable) versus those over which they have some control (mutable). We propose that brands should use this classification system to understand when they might be more or less at risk, due to different types of brand safety incidents. For example, when they lack much control over the moderator—such as their predominant offering (service- vs. product-oriented), brand type (utilitarian or hedonic), consumers' perceptions of different categories of unsafe topics, or the type of negative adjacent content—brands need to develop risk

assessments that specify different digital advertising strategies. The extent to which brands seek to mitigate brand safety risks due to factors outside of their control is a risk-management decision and our findings suggest that brand safety in digital advertising channels is a valid type of risk for companies to consider, particularly if they use digital advertising a lot.

Our Study 1 results imply, while fully acknowledging their descriptive nature, that managers of service-oriented and hedonic brands must be particularly vigilant in monitoring digital environments for brand safety incidents, because their brands are at risk of experiencing significantly greater negative effects due to safety incidents. Such strategies also should account for the brand's image and target consumers' behaviors and perceptions (which also are typically immutable classifications, assuming no major rebranding). If consumers perceive a higher fit between the brand's image and the unsafe content or do not consider the adjacent content particularly unsafe (e.g., men or more conservative individuals in Studies 2c and 4a; someone who accepts cannabis use; Study 4b), the brand outcomes could be neutral or positive rather than negative. On the other side, our experimental findings show that there may be some backfiring effects for consumers who are very connected to, or committed to, specific types of brands, in line with prior research (Mosley, Schweidel, and Zhang 2024), and possibly in line with Study 1 (as this effect only appeared for Disney connections which is a more service-oriented and hedonic brand than some others examined in our research; Web Appendix E).

Thus, accounting for the brand's image and target audience should define how the brand deals with adjacency to different types of "unsafe" digital content, with the recognition that in some scenarios, there is no managerial imperative to combat the relatively inconsequential risk. For other types of immutable factors though, the recommendations differ. Brands typically have little control over the type of content that appears adjacent to their advertising, but some

incidents sparked by negative content threaten negative effects for all brands (e.g., hate speech). It is therefore imperative, from a risk assessment perspective, to monitor the digital and social media landscape to anticipate different types of brand safety incidents and plan response strategies that can be deployed in a timely fashion.

Whereas prior brand safety research highlights the effects of attributions of responsibility (Manatt, Avital, and Ofer 2018), perceived brand responsibility does not drive the effects in our studies, nor does it consistently moderate our findings. In turn, we argue that brands should establish well-resourced teams to monitor and manage safety incidents, as well as develop pertinent response strategies, rather than devoting effort to assigning responsibility and blame to external platforms or algorithms in response to incidents. Regardless of their level of understanding about how online ads get placed and their blame attributions, consumers exhibit diminished brand outcomes after brand safety incidents. Therefore, brands need to develop more reactive, non-recriminatory approaches. To offer some initial support for this assertion, we show experimentally that the timing of a public relations apology, without assigning blame, after a brand safety incident significantly affects how well consumers receive the statement and their overall brand evaluations (Web Appendix O). In turn, we reemphasize the need for brands to engage in careful monitoring, because if they can respond to an incident quickly, it might mitigate future negative outcomes (and possibly even engender positive associations). We hope additional research will take up the question of how the brand's chosen response, in tandem with the timing of that response, influences various brand-relevant outcomes.

Turning instead to more mutable moderators (e.g., familiarity, liking, self–brand connection), we recommend that brands develop proactive programs, in addition to their reactive strategies. The evidence from Study 1 (while again noting the descriptive nature of these results)

implies brands should proactively invest in programs that build up the stock of the brand with respect to these variables as it can potentially help shield it from the worst harms of brand safety incidents. For example, engaging in expanded, public corporate social responsibility efforts might increase consumers' general familiarity with or liking of the brand, prior to any potential safety incidents. Improving liking, familiarity, and self–brand connections generally require longer-term, proactive investments, though brands also have options for encouraging more positive consumer perceptions after an incident, regardless of their previous perceptions. Due to the results from Study 1, in an exploratory study (Web Appendix P), we postulate that brands can leverage language that cites and evokes greater brand liking or familiarity in digital ads, run after a brand safety incident. These initial findings suggest that consumers, even without any preconceived notions of brands, offer more positive brand evaluations after receiving ads that include such terms (cf. a control ad).

Finally, we hope continued research will explore other potential antecedents of brand safety incidents, to give brands additional tools and measures to include in their risk assessments. For example, some factors might indicate if a brand safety incident is likely to spark a critical mass of media attention or online firestorm (Herhausen et al. 2019), which can help brands prepare to engage in timely public relations efforts. We speculate that the frequency of media coverage, the nature of the incident (e.g., type of unsafe content, frequency of prior incidents), or how media outlets report on incidents (e.g., citing the role of platforms, algorithms, or brands) all might represent important influences, and we leave it to further research to specify their effects.

References

Ahluwalia, Rohini, H. Rao Unnava, and Robert E. Burnkrant (2001), "The Moderating Role of Commitment on the Spillover Effect of Marketing Communications," *Journal of Marketing Research,* 38 (4), 458-470.

Aribarg, Anocha, and Eric M. Schwartz (2020), "Native Advertising in Online News: Trade-offs Among Clicks, Brand Recognition, and Website Trustworthiness," *Journal of Marketing Research*, 57 (1), 20-34.

Bart, Yakov, Venkatesh Shankar, Fareena Sultan, and Glen L. Urban (2005), "Are the Drivers and Role of Online Trust the Same for All Web Sites and Consumers? A Large Scale Exploratory Empirical Study," *Journal of Marketing*, 69 (October), 133–52.

Belanche, Daniel, Carlos Flavian, and Alfredo Perez-Rueda (2017), "Understanding Interactive Online Advertising: Congruence and Product Involvement in Highly and Lowly Arousing, Skippable Video Ads." *Journal of Interactive Marketing*, 37, 75–88.

Bellman, Steven, Ziad HS Abdelmoety, Jamie Murphy, Shruthi Arismendez, and Duane Varan (2018), "Brand Safety: The Effects of Controversial Video Content on Pre-roll Advertising," *Heliyon*, 4 (12).

Borah, Abhishek, and Gerard J. Tellis (2016), "Halo (Spillover) Effects in Social Media: Do Product Recalls of One Brand Hurt or Help Rival Brands?" *Journal of Marketing Research*, 53 (2), 143-160.

Buechel, Eva C., and Chris Janiszewski (2014), "A Lot of Work or a Work of Art: How the Structure of a Customized Assembly Task Determines the Utility Derived from Assembly Effort," *Journal of Consumer Research*, 40 (5), 960-972.

Bushman, Brad J (2005), "Violence and Sex in Television Programs Do Not Sell Products in Advertisements," *Psychological Science*, 16 (9), 702-708.

——— (2007), "That Was a Great Commercial, but What Were They Selling? Effects of Violence and Sex on Memory for Products in Television commercials," *Journal of Applied Social Psychology*, 37 (8), 1784-1796.

Chaudhuri, Arjun, and Morris B. Holbrook (2001), "The Chain of Effects from Brand Trust and Brand Affect to Brand Performance: The Role of Brand Loyalty," *Journal of Marketing*, 65 (2), 81-93.

Chen, Charlene Y., Leonard Lee, and Andy J. Yap (2017), "Control Deprivation Motivates Acquisition of Utilitarian Products," *Journal of Consumer Research*, 43 (6), 1031-1047.

Chung, Jaeyeon, Leonard Lee, Donald R. Lehmann, and Claire I. Tsai (2023), "Spending Windfall ('Found') Time on Hedonic versus Utilitarian Activities," *Journal of Consumer Research*, 49 (6), 1118-1139.

Cleeren, Kathleen, Harald J. Van Heerde, and Marnik G. Dekimpe (2013), "Rising from the Ashes: How Brands and Categories can Overcome Product-harm Crises," *Journal of Marketing*, 77 (2), 58-77.

Culotta, Aron, and Jennifer Cutler (2016), "Mining Brand Perceptions from Twitter Social Networks," *Marketing Science*, 35 (3), 343-362.

Dai, Hengchen, Cindy Chan, Cassie Mogilner (2020), "People Rely Less on Consumer Reviews for Experiential than Material Purchases," *Journal of Consumer Research*, 46 (6), 1052–1075.

Dawar, Niraj, and Madan M. Pillutla (2000), "Impact of Product-Harm Crises on Brand Equity: The Moderating Role of Consumer Expectations," *Journal of Marketing Research*, 37 (2), 215-226.

Doney, Patricia M., and Joseph P. Cannon (1997), "An Examination of the Nature of Trust in Buyer–seller Relationships," *Journal of Marketing*, 61 (2), 35-51.

Escalas, Jennifer Edson, and James R. Bettman (2003), "You Are What They Eat: The Influence of Reference Groups on Consumers' Connections to Brands," *Journal of Consumer Psychology*, 13 (3), 339-348.

Gallup (2023), "Confidence in Institutions Dips," https://tinyurl.com/2bs3k535.

Garbarino, Ellen, and Mark S. Johnson (1999), "The Different Roles of Satisfaction, Trust, and Commitment in Customer Relationships," *Journal of Marketing*, 63 (2), 70-87.

Graham, Megan (2023), "Media Veteran Lou Paskalis Joins Group to Encourage Brands to Advertise on News," *The Wall Street Journal*, accessed January 15, 2024 at: http://tinyurl.com/ysta6bmx.

Grégoire, Yany, Thomas M. Tripp, and Renaud Legoux (2009), "When Customer Love Turns into Lasting Hate: The Effects of Relationship Strength and Time on Customer Revenge and Avoidance," *Journal of Marketing*, 73 (6), 18-32.

Hansen, Nele, Ann-Kristin Kupfer, and Thorsten Hennig-Thurau (2018), "Brand Crises in the Digital Age: The Short-and Long-term Effects of Social Media Firestorms on Consumers and Brands," *International Journal of Research in Marketing*, 35 (4), 557-574.

Hasford, Jonathan, David M. Hardesty, and Blair Kidwell (2015), "More Than a Feeling: Emotional Contagion Effects in Persuasive Communication." *Journal of Marketing Research*, 52 (6), 836-847.

Hayes, Andrew F. (2017), "Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-based Approach," New York, NY: Guilford Press.

Herhausen, Dennis, Stephan Ludwig, Dhruv Grewal, Jochen Wulf, and Marcus Schoegel (2019), "Detecting, Preventing, and Mitigating Online Firestorms in Brand Communities," *Journal of Marketing*, 83 (3), 1-21.

Holbrook, Morris B., and Elizabeth C. Hirschman (1982), "The Experiential Aspects of Consumption: Consumer Fantasies, Feelings, and Fun," *Journal of Consumer Research*, 9 (2), 132-140.

Hotz-Behofsits, Christian, Nils Wlömert, and Nadia Abou Nabout (2025), "Natural Affect DEtection (NADE): Using Emojis to Infer Emotions from Text," *Journal of Marketing*, 0(ja). https://doi.org/10.1177/00222429251315088

Hsu, Tiffany (2020), "Cruise Line Ads Get Caught in a Coronavirus News Cycle," *The New York Times*, January 7. https://tinyurl.com/48dz352b

Huh, Jisu, and Leonard N. Reid (2007), "Do Consumers Believe Advertising is Negatively Affected when Placed Near News Perceived as Biased?" *Journal of Current Issues & Research in Advertising*, 29 (2), 15-26.

Hutto, Clayton, and Eric Gilbert (2014), "Vader: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," In *Proceedings of the International AAAI Conference on Web and Social Media*, 8 (1), 216-225.

Interactive Advertising Bureau (2018), "Brand Safety in Today's Digital Context." https://www.iabsa.net/assets/Usedebbieiabsanet/IAB_SA_Brand_Safety_White_Paper,_25_Oct_2018.pdf

Johar, Gita V., Matthias M. Birk, and Sabine A. Einwiller (2010), "How to Save Your Brand in the Face of Crisis," *MIT Sloan Management Review*, 51 (4), 57-64.

Johnson, Ross W., Clay Voorhees, and Farnoosh Khodakarami (2023), "Is Your Brand Protected?: Assessing Brand Safety Risks In Digital Campaigns," *Journal of Advertising Research*, 63 (3), 205-220.

Keller, Kevin Lane (1993), "Conceptualizing, Measuring, and Managing Customer-Based Brand Equity," *Journal of Marketing*, 57 (1), 1–22.
Keller, Kevin L. (2001), "Building Customer-based Brand Equity: A Blue Print for Creating Strong Brands."
Keller, Kevin Lane, and Donald R. Lehmann (2006), "Brands and Branding: Research Findings and Future Priorities," *Marketing Science,* 25 (6), 740-759.
Khamitov, Mansur, Koushyar Rajavi, Der-Wei Huang, and Yuly Hong (2024), "Consumer Trust: Meta-analysis of 50 Years of Empirical Research," *Journal of Consumer Research,* 51 (1), 7-18.
Kim, Laura R., and Nancy S. Kim (2011), "A Proximity Effect in Adults' Contamination Intuitions," *Judgment and Decision Making*, 6 (3), 222-229.
Knox, George, and Rutger Van Oest (2014), "Customer Complaints and Recovery Effectiveness: A Customer Base Approach," *Journal of Marketing*, 78 (5), 42-57.
Koschate-Fischer, Nicole, Wayne D. Hoyer, and Christiane Wolframm (2019), "What if Something Unexpected Happens to my Brand? Spillover Effects from Positive and Negative Events in a Co-branding Partnership," *Psychology & Marketing*, 36 (8), 758-772.
Kronrod, Ann, and Shai Danziger (2013), ""Wii Will Rock You!" The Use and Effect of Figurative Language in Consumer Reviews of Hedonic and Utilitarian Consumption," *Journal of Consumer Research,* 40 (4), 726-739.
Lambrecht, Anja, and Catherine Tucker (2013), "When Does Retargeting Work? Information Specificity in Online Advertising," *Journal of Marketing research*, 50 (5), 561-576.
Lee, Chunsik, Junga Kim, and Joon Soo Lim (2021), "Spillover Effects of Brand Safety Violations in Social Media," *Journal of Current Issues & Research in Advertising*, 42 (4), 354-371.
Lei, Jing, Niraj Dawar, and Jos Lemmink (2008), "Negative Spillover in Brand Portfolios: Exploring the Antecedents of Asymmetric Effects," *Journal of Marketing*, 72 (3), 111-123.
Li, Jingjing, Ahmed Abbasi, Amar Cheema, and Linda B. Abraham (2020), "Path to Purpose? How Online Customer Journeys Differ for Hedonic versus Utilitarian Purchases," *Journal of Marketing,* 84 (4), 127-146.
Liu, Yan, and Venkatesh Shankar (2015), "The Dynamic Impact of Product-harm Crises on Brand Preference and Advertising Effectiveness: An Empirical Analysis of the Automobile Industry," *Management Science*, 61 (10), 2514-2535.
Manatt, Kara, Daniel Avital, and Ben Ofer (2018), "The Brand Safety Effect: How Unsafe Ad Placement Impacts Consumer Brand Perception," MAGNA website.
McKnight, D. Harrison, Vivek Choudhury, and Charles Kacmar (2002), "The Impact of Initial Consumer Trust on Intentions to Transact with a Web Site: A Trust Building Model," *The Journal of Strategic Information Systems*, 11 (3-4), 297-323.
Mizerski, Richard W. (1982), "An Attribution Explanation of the Disproportionate Influence of Unfavorable Information," *Journal of Consumer Research*, 9 (3), 301-310.
Moorman, Christine, Gerald Zaltman, and Rohit Deshpande (1992), "Relationships between Providers and Users of Market Research: The Dynamics of Trust Within and Between Organizations," *Journal of Marketing Research*, 29 (3), 314-328.
Morgan, Robert M. and Shelby D. Hunt (1994), "The Commitment-Trust Theory of Relationship Marketing," *Journal of Marketing*, 58 (July), 20-38.

Mosley, Buffy, David A Schweidel, and Kunpeng Zhang (2024), "When Connection Turns to Anger: How Consumer-Brand Relationship and Crisis Type Moderate Language on Social Media," *Journal of Consumer Research*, 50 (5), 907-922.

Naylor, Rebecca Walker, Cait Poynor Lamberton, and David A. Norton (2011), "Seeing Ourselves in Others: Reviewer Ambiguity, Egocentric Anchoring, and Persuasion," *Journal of Marketing Research*, 48 (3), 617-631.

Nian, Tingting, Yuheng Hu, and Cheng Chen (2021), "Examining the Impact of Television-Program-Induced Emotions on Online Word-of-Mouth toward Television Advertising," *Information Systems Research*, 32 (2), 605-632.

Patterson, Paul G., Elizabeth Cowley, and Kriengsin Prasongsukarn (2006), "Service Failure Recovery: The Moderating Impact of Individual-level Cultural Value Orientation on Perceptions of Justice," *International Journal of Research in Marketing*, 23 (3), 263-277.

Pham, Michel Tuan (1998), "Representativeness, Relevance, and the Use of Feelings in Decision Making," *Journal of Consumer Research*, 25 (2), 144–159.

——— (2013), "The Seven Sins of Consumer Psychology," *Journal of Consumer Psychology*, 23 (4), 411-423.

Power, John, Susan Whelan, and Gary Davies (2008), "The Attractiveness and Connectedness of Ruthless Brands: The Role of Trust," *European Journal of Marketing*, 42 (5/6), 586-602.

Roehm, Michelle L., and Alice M. Tybout (2006), "When Will a Brand Scandal Spill Over, and How Should Competitors Respond?" *Journal of Marketing Research*, 43 (3), 366-373.

Rust, Roland T., William Rand, Ming-Hui Huang, Andrew T. Stephen, Gillian Brooks, and Timur Chabuk (2021), "Real-Time Brand Reputation Tracking Using Social Media," *Journal of Marketing*, 85 (4), 21-43.

Sahni, Navdeep S., and Harikesh S. Nair. (2020), "Sponsorship Disclosure and Consumer Deception: Experimental Evidence from Native Advertising in Mobile Search," *Marketing Science*, 39 (1), 5-32.

Smith, Amy K., Ruth N. Bolton, and Janet Wagner (1999), "A Model of Customer Satisfaction with Service Encounters Involving Failure and Recovery," *Journal of Marketing Research*, 36 (3), 356-372.

Srinivasan, Raji, and Gülen Sarial-Abi (2021), "When Algorithms Fail: Consumers' Responses to Brand Harm Crises Caused by Algorithm Errors," *Journal of Marketing*, 85 (5), 74-91.

Votolato, Nicole L., and H. Rao Unnava (2006), "Spillover of Negative Information on Brand Alliances," *Journal of Consumer Psychology*, 16 (2), 196-202.

Whelan, Jodie, and Niraj Dawar (2016), "Attributions of Blame Following a Product-harm Crisis Depend on Consumers' Attachment Styles," *Marketing Letters*, 27, 285-294.