From brand safety to suitability: advertisers in platform governance

Rachel Griffin

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 12 **Issue** 3

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

From brand safety to suitability: advertisers in platform governance

Rachel Griffin *Paris Institute of Political Studies* rachel.griffin@sciencespo.fr

DOI: https://doi.org/10.14763/2023.3.1716

Published: 11 July 2023
Received: 14 October 2022 **Accepted:** 25 April 2023

Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Griffin, R. (2023). From brand safety to suitability: advertisers in platform governance. *Internet Policy Review*, *12*(3). https://doi.org/10.14763/2023.3.1716

Keywords: Platform governance, Social media, Platform regulation, Content regulation, Digital Service Act

Abstract: Scholarship has long identified the business imperative to create an advertiser-friendly environment as a key influence on social media content moderation. However, "brand safety" – the industry term for advertisers' measures to avoid content perceived as reflecting negatively on their brands – remains understudied. Drawing on policy statements from industry actors, as well as extant academic literature, this article makes four contributions. First, it proposes four distinct mechanisms through which branding imperatives influence platforms' content governance. Second, it highlights two current trends: growing efforts by major advertisers to directly influence platforms' content policies, and a shift in industry terminology from brand safety (avoiding content widely considered objectionable) to "suitability" (evaluating appropriate content for a particular brand) – which promises advertisers greater customisation, but in fact promotes the standardisation of content governance across major platforms. Third, it explores the policy implications of these developments, in particular for equal participation and freedom of public debate on social media. Finally, it briefly explores the relevance to these concerns of the EU's 2022 Digital Services Act, suggesting that it fails to adequately address a marketised logic in which the production and distribution of online media content is increasingly shaped by what is deemed suitable for branding objectives.

Introduction

Despite their diverse features and audiences, the most popular social media platforms in the West today share a business model: targeted advertising. Scholarship has identified the need to attract advertisers as a key influence on platforms' content moderation (Klonick, 2018; Gillespie, 2018). However, the precise mechanisms through which advertisers influence platform governance remain understudied.

In the advertising industry, *brand safety* refers to "ensuring that a brand's ad should not appear adjacent to content or in a context that can damage the advertised brand" (Lee et al., 2021). To this end, as this article describes, corporate advertisers and industry associations have demanded various technical and policy changes from platforms. Advertisers' commercial imperatives drive selective censorship of user content and homogenisation of platform policies across the industry. Yet, while brand safety tools and strategies have attracted some attention in media and cultural studies (Kumar, 2019; Craig & Cunningham, 2019; Bishop, 2021), economics (Madio & Quinn, 2023) and marketing industry journals (Marvin & Meisel, 2017; Lee et al., 2021), they have mostly gone unmentioned in platform governance literature.

To remedy this, this article makes four contributions. First, it proposes a typology of four mechanisms through which branding imperatives influence content governance. Second, it highlights two current trends: intensifying efforts by advertisers to exert influence, and a shift in industry terminology from brand safety (avoiding content widely considered objectionable) to *suitability* (evaluating appropriateness for particular brands). Third, the article discusses these developments' policy implications, contextualising them in the longer history of advertiser influence in media production. Finally, it briefly explores how these policy concerns are addressed by the EU's 2022 Digital Services Act (DSA).

It concludes that the increasingly prominent concept of brand suitability can offer insights into how advertising shapes social media governance. While promising advertisers greater customisation, brand suitability tools and policies are in fact highly standardised and likely to promote further homogenisation across major platforms, as well as expanding and institutionalising advertiser influence. In contrast to safety, which focuses attention on the management of discrete risks, suitability evokes broader logics of content moderation and curation, with only content considered suitable for major advertisers permitted to become widely visible. In this respect, brand suitability offers a useful lens for closer scholarly engagement with the industry-wide market forces shaping content curation, online visibil-

ity and platform governance.

Brand safety and content governance

This article presents the first systematic study of how brand safety efforts influence social media governance, although existing literature has explored specific aspects of brand safety. The most detailed study comes from Bishop (2021), who analyses third-party "influencer management tools" for brands partnering with social media creators, and highlights how commercial imperatives and discriminatory assumptions guide brand safety assessments. However, Bishop does not directly address the brand safety policies and tools offered by platforms themselves – which have broader impacts, as they extend to all users and content, not only specific influencers partnering with brands. Other media scholars have examined changes in YouTube's policies following the 2017 "Adpocalypse", a scandal revolving around the placement of ads alongside terrorism-related content, which led major advertisers to boycott YouTube, demanding enhanced brand safety measures (Kumar, 2019; Craig & Cunningham, 2019; Caplan & Gillespie, 2020). However, there is little research into brand safety tools at other major platforms.

This section thus aims to build on this existing literature with a broader overview of brand safety tools and policies at leading platforms, based on an examination of relevant policy statements from Meta, YouTube and TikTok. It identifies four main mechanisms through which brand safety concerns influence content governance: generally-applicable moderation policies prohibiting unsafe content; generally-applicable demonetisation policies which exclude content from ad placements; customisable advertiser settings; and indirect influence on recommendations and platform design.

Generally-applicable moderation policies

Although Article 3(t) DSA includes demonetisation (removal of ads) and demotion in recommendations in its definition of moderation, and some scholars define it even more broadly to include all platform decisions influencing the production and circulation of user-generated content (Gillespie, 2022; Douek, 2022), for clarity this section will focus only on moderation policies setting out what users are allowed to post on a platform. Demonetisation and recommendations involve distinct incentives for advertisers and creators and will be addressed in the following sections.

Scholarship on these policies identifies the need to ensure that users are not put

off by explicit or upsetting content, and adverts are not associated with offence or controversy, as a primary reason for their development and implementation (Klonick, 2018; Gillespie, 2018; Roberts, 2018). For example, strict bans on sexual content (currently in place at all major Western platforms except Twitter: Bayley, 2021) are seen as significantly influenced by advertisers' aversion to content not considered "family-friendly" (Roberts, 2018; Are & Paasonen, 2021). At the same time, these bans are often selectively enforced, such that highly sexualised imagery in adverts themselves, and commodifiable "mainstream" content like celebrity posts which are valuable vehicles for advertising, frequently escapes moderation (Roberts, 2018; Are & Paasonen, 2021; Waldman, 2021).

Major platforms openly state that these policies are influenced by advertiser interests. For example, Meta and TikTok's brand safety pages for advertisers repeatedly mention moderation policies as the primary safeguard against brand safety violations, framing other tools as optional extra precautions (Meta, n.d.a, n.d.b, n.d.c; TikTok, n.d.). Notably, Meta (n.d.c) claims that Facebook's content policies uniformly match or exceed the "brand safety floor" defined by the Global Alliance for Responsible Media (GARM), an industry initiative from the World Federation of Advertisers (WFA, n.d.a). That is, nothing the industry considers broadly brand unsafe is allowed on Facebook at all. As section 4 will show, given discrepancies between what corporate advertisers consider safe and what the public might consider interesting or important, such a statement has major implications for online media freedom.

Advertisers have also openly sought to influence moderation. During 2020's Black Lives Matter protests, numerous large advertisers announced a boycott of Facebook ads until it improved its hate speech policies.[1] Shortly afterwards, major platforms formally agreed to align their definitions of prohibited hate speech with the WFA's brand safety definitions (WFA, 2020). The 2020 boycott and other similar campaigns were encouraged by NGOs including Sleeping Giants and Color of Change, who consider this a useful lever to incentivise more action by platforms against racism (Wodinsky, 2020; Hendrix, 2021). It is questionable, however, whether the interests of major corporate advertisers can generally be expected to coincide with the demands of racial and social justice movements.

1. These boycotts were in practice fairly limited, and affected Meta more through bad publicity than financial losses (Wodinsky, 2020). However, as section 3 discusses, a general trend towards declining advertising revenue may increase advertisers' financial leverage over platforms.

Generally-applicable (de)monetisation policies

Demonetisation is an alternative moderation measure, where platforms continue hosting content but do not accompany it with ads, alleviating brand safety concerns. YouTube's demonetisation policies have attracted particular attention, because it was historically the only major platform directly sharing ad revenue with creators (Caplan & Gillespie, 2020); for a minority of successful creators, this can be a primary income source (Glatt, 2022). However, Facebook also now allows revenue-sharing for certain content on public pages. In these cases, demonetisation directly penalises creators, significantly influencing how they create and present content (Craig & Cunningham, 2019; Glatt & Banet-Weiser, 2021).

However, even where ad revenue is not shared and demonetisation does not directly affect creators, it can have indirect impacts. Since platforms' recommendation algorithms are designed to maximise ad revenue, their incentives would logically be to favour content running with ads (Kumar, 2019; Glatt & Banet-Weiser, 2021). Although YouTube has denied that demonetisation affects recommendations (Creator Insider, 2018), investigations by YouTubers' Union founder JoergSprave (2018) suggest that the two are, in his case, closely correlated (Kumar, 2019). It is possible that platforms do not directly use demonetisation as a variable in recommendation algorithms, but do design them to deprioritise the same types of content that are likely to be demonetised – which would make obvious business sense. Importantly, recommendations do not affect the visibility of already-created content, but what creators produce in the first place, as they strategically adapt to recommendation algorithms' criteria (Cotter, 2019; Bishop, 2018; Glatt, 2022).

Since the 2017 boycott, YouTube automatically analyses the brand safety of all videos eligible for monetisation (which requires a certain number of views: Kumar, 2019). To facilitate this, creators are asked to add thematic tags (Rodriguez, 2022). Demonetisation can also occur later: where users report a video, moderators may decide it does not violate moderation policies but is not brand safe, and demonetise instead of deleting it (Rodriguez, 2022). Content deemed inappropriate for children and blocked for under-18s is generally also demonetised (Rodriguez, 2022).

Notably, YouTube states that "content discussing terrorism or sensitive current events like war, death, or tragedy" is always unsafe, and will be demonetised (Google, n.d.) – a strong disincentive for professional creators to discuss many political topics. Meta's policy is even broader: anything involving "suggestive lan-

guage", "discussion" of nudity or "revealing" clothing may be demonetised. It is also extremely vague, only ever stating what "may" affect monetisation (Meta, n.d.d). With little clear guidance, creators hoping to monetise their content are incentivised to err on the side of self-censorship.

Tailored advertiser tools

Alongside these generally-applicable policies, YouTube, Meta and TikTok offer optional tools for individual advertisers to control ad placements: for example, excluding certain URLs or types of content (such as livestreams: Meta, n.d.a). The most prominent and easiest-to-use (indeed, they are enabled by default) are *inventory controls*. These settings limit the inventory of available content beyond what has already been demonetised, based on tiered classifications of content as suitable for advertisers with high, medium or low risk tolerance.

YouTube introduced "expanded" (high-risk), "standard" (medium) and "limited" (low-risk) inventory settings following the 2017 boycott (Google, n.d.; Kumar, 2019). Meta (2019, n.d.e) introduced "full", "standard" and "limited" inventory settings for videos in 2019, and now offers "expanded", "moderate" and "limited" inventories for all Facebook and Instagram posts (Meta, 2022, n.d.e). TikTok (n.d.) offers similar "full", "standard" and "limited" tiers. While little detail is available about how these classifications are operationalised, they appear to rely on AI analysis of content, as well as metadata such as video titles. To implement inventory filters, Meta and TikTok both partnered with brand safety software firm Zefr, which claims to combine machine learning with targeted human review (Zefr, n.d.a, n.d.b).

In most cases, advertisers are opted into the middle "standard" tier by default (Meta does this for in-content video ads, but defaults to "expanded" inventory for other content: Meta, n.d.e). The exclusion by many or most advertisers of content excluded from "standard" inventory may therefore have similar effects to complete demonetisation: creators will be disincentivised from creating such content, and platforms from recommending it.

Recommendations and platform design

Although conclusive evidence that demonetisation and inventory exclusion affect recommendations is lacking, this is one obvious way that brand safety might affect the kinds of content platforms choose to promote. However, given platforms' economic reliance on advertisers, they are also generally incentivised to encourage users, through recommendations and other design choices, to create the kinds of content that advertisers favour (Carah & Brodmerkel, 2022). Historically, this has

meant content considered upbeat, uncontroversial and family-friendly (Baker, 1992; Shepherd, 2014; Bishop, 2018, 2021).

While teasing out the exact reasons behind companies' decisions is difficult, media scholars have identified some design decisions that seem aligned with these incentives. For example, 'like' buttons encourage positive and easily-quantifiable interactions (Shepherd, 2014) – suitable both for creating a "buying mood" (Baker, 1992, pp. 2153-55) and for algorithmically analysing user behaviour. Bishop (2018) argues that YouTube disproportionately recommends creators who conform to dominant gender and class norms, linking this to advertisers' demands for gender-segregated audiences. Platforms have also been accused of promoting unattainable and exclusionary beauty standards via features like facial "enhancement" filters – which arguably makes business sense insofar as it creates a favourable environment for beauty, fashion, diet and "wellness" adverts (Griffin, 2023a).

Current trends

Against this backdrop, two ongoing trends can be identified. First, advertisers appear to be intensifying efforts to directly influence platform governance. In general, brand safety has become a bigger industry concern since the rise of programmatic advertising (Marvin & Meisel, 2017). More recently, as platform governance has attracted more public attention and advertisers have been publicly criticised for funding harmful content, advertisers have successfully demanded policy changes and expanded brand safety tools, as section 2 described.

Some factors suggest advertiser influence may continue to increase. Advertising revenue growth at major platforms is weak, which analysts have ascribed in varying degrees to both macroeconomic and industry-specific factors (Milmo, 2023; Seufert, 2023). Greater competition between platforms to attract advertisers could increase their leverage. Perhaps reflecting this, major platforms increasingly coordinate with the advertising industry, particularly through GARM. As noted above, Meta (n.d.c) claims all Facebook content policies comply with GARM's "brand safety floor". Zefr's content classification system – the basis for Meta and TikTok's inventory controls – is based on GARM's brand suitability taxonomy (Zefr, n.d.a, n.d.b). At the same time, platforms are also increasingly experimenting with revenue sources other than advertising, notably subscriptions (Milmo, 2023) and e-commerce (Goanta, 2023). Should e-commerce become more significant, it will also create pressures for platforms to moderate and curate content in ways that suit branding and marketing goals.

This points to the second notable trend: a terminological shift from safety, generally understood as addressing clearly inappropriate content, to suitability, denoting a more customised assessment of appropriateness for a particular advertiser (Zefr, n.d.c; n.d; DoubleVerify, 2021).

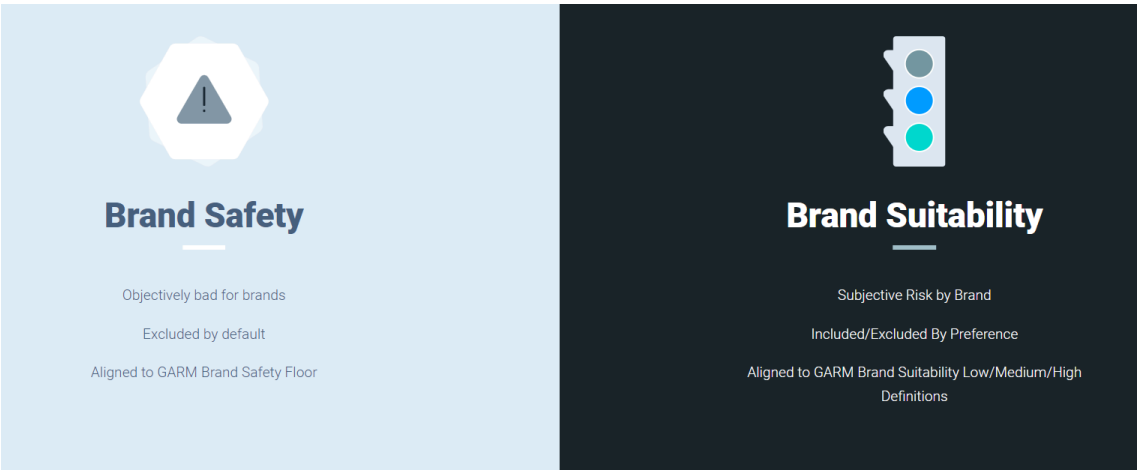

FIGURE 1: (Zefr, n.d.c).

Despite industry rhetoric around customisation, brand suitability tools are highly standardised. Most advertisers will likely use default inventory tiers, which are similar across platforms, as they all draw on GARM's brand suitability framework.[2] This taxonomy, which defines low-, medium- and high-risk content across 11 topics, is highly formulaic, as the excerpt in Figure 2 illustrates. Many categories do not appear individually thought through, with identical phrasing recurring across topics.

2. YouTube does not explicitly say its brand suitability classifications are based on GARM's framework, but does cooperate closely with GARM (Wolinsky, 2022).

Online piracy	• Glamorization /Gratuitous depiction of Online Piracy	• Dramatic depiction of Online Piracy presented in the context of entertainment • Breaking News or Op-Ed coverage of Online Piracy	• Educational, Informative, Scientific treatment of Online Piracy • News feature stories on the subject
Hate speech & acts of aggression	• Depiction or portrayal of hateful, denigrating, or inciting content focused on race, ethnicity, gender, sexual orientation, gender identity, age, ability, nationality, religion, caste, victims and survivors of violent acts and their kin, immigration status or serious disease sufferers, in a non-educational, informational, or scientific context	• Dramatic depiction of hate speech/acts presented in the context of entertainment • Breaking News or Op-Ed coverage of hate speech/acts	• Educational, Informative, Scientific treatment of Hate Speech • News features on the subject
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	• Glamorization /Gratuitous depiction of profanity and obscenity	• Dramatic depiction of profanity and obscenities presented in the context of entertainment by genre • Breaking News or Op-Ed coverage of profanity and obscenities Genre based use of profanity, gestures, and other actions that may be strong, but might be expected as generally accepted language and behavior	• Educational or Informative, treatment of Obscenity or Profanity • News feature stories on the subject

FIGURE 2: The left, middle and right columns respectively describe high-, medium- and low-risk content (GARM, 2022).

Both trends will promote more uniformity in content governance. Platforms face similar economic pressures from advertisers, and respond with similar tools – often directly copying each other (Gillespie, 2018), as with Meta and TikTok's introduction of inventory controls modelled on YouTube's. Common reliance on GARM's framework and on software-as-a-service providers like Zefr – which aims to provide "industry-standard" classifications (Zefr, n.d.d) – also drives homogenisation. Zefr has also developed a dataset of content examples for GARM's categories, presented as (prospective) industry-standard training data for AI classifiers (Morra, 2021). As standardised norms and standard technical tools for operationalising them continue to be developed and institutionalised, the same content is likely to be deemed unsafe or unsuitable across most major platforms.

Brand suitability also denotes a small but significant shift in emphasis. While safety tends to evoke the elimination or management of specific dangers (Thomasen, 2023), suitability suggests more emphasis on what positively suits advertisers' business needs. In this sense, this terminological shift reflects the ongoing expansion and institutionalisation of advertiser influence – which, as section 2(d) argued, not only shapes moderation but also content curation and governance more broadly. Thus, the shift to brand suitability also aligns with the increasing focus on modulating visibility as a means of content governance (Zeng & Kaye, 2021) as personalised algorithmically-curated feeds become increasingly central to user experiences (Riemer & Peter, 2021; Narayanan, 2023). Given platforms' incentives to attract advertisers and maximise the effectiveness of adverts, brand suitability considerations can be expected to influence what they positively recommend in

these feeds, as well as what they eliminate. Epitomising this shift, TikTok (2022) recently announced that it would reward aspiring influencers by recommending their content more if they promoted specified "brand missions". Moving beyond brand safety policies which focus on identifying and excluding damaging content, platforms now also openly promote the creation of content tailored to suit brands.

Policy implications

The mechanisms identified in section 2 and the increasing influence of brand safety and suitability in content governance have concerning implications for equal and open participation in social media. In some ways, this represents a continuation of trends that pre-existed digital media. Critical political economy of the media literature has long highlighted how publishers' structural dependence on advertisers allows them to shape media production and distribution to their own advantage, often acting as "the most consistent and the most pernicious 'censors' of media content" (Baker, 1992, p. 2009) – from pressuring publishers not to cover specific topics to explicitly demanding content that is pro-capitalist, uncontroversial and creates a "buying mood" (Baker, 1992, pp. 2153-55; Herman & Chomsky, 1995). Gloria Steinem (1990) memorably describes how *Ms* magazine struggled financially because it refused to provide superficial, pro-consumerist "complementary copy" to beauty adverts. As well as disfavouring political content and critical journalism, what is brand safe has always been gendered, raced and heteronormative, and contemporary brand safety standards show significant continuities with these historical dynamics.

Discrimination

Bishop (2021) shows that third-party brand safety tools enforce gendered, heteronormative standards of sexual morality and associate racialised creators with risk. Platforms' own tools exhibit similar tendencies. Notoriously, YouTube has been shown to indiscriminately demonetise LGBTQIA+ content (Kumar, 2017; Rodriguez, 2022): an independent investigation found that keywords like "gay" triggered demonetisation, while identical videos with only those words changed were permitted (Alexander, 2019). As section 2(b) outlined, this directly influences creators' behaviour. Some do not tag their content with LGBTQIA+-related keywords to avoid demonetisation, making it less accessible to audiences (Glatt & Banet-Weiser, 2021). Demonetisation may also reduce recommendations of LGBTQIA+ content, diminishing its overall visibility (Kumar, 2019).

YouTube apologised in 2017 for some obvious instances of discrimination, but

framed these as individual technical errors, without accepting criticism of the underlying system which deems sex and sexuality inappropriate. Revealingly, its examples of erroneously demonetised videos – a lesbian wedding, a boy coming out to his grandmother – involve "homonormative" images of queer people as desexualised and assimilated into traditional family institutions (Southerton et al., 2020). Implicitly, queer sexuality and non-normative representations of queerness remain "unsafe" and will be penalised.

News and politics

As well as LGBTQIA+ issues, advertisers' demands for positive, broadly uncontroversial content affect political speech and activism more broadly. YouTube demonetises all discussions of "terrorism or sensitive current events like war, death, or tragedy", even for advertisers selecting "expanded inventory" (Google, n.d.). (Semi-)professional creators are thus penalised for discussing many major current events; YouTubers focusing on news and politics have stated that this discourages them from addressing serious topics (Craig & Cunningham, 2019). The GARM taxonomy on which Meta and TikTok's inventory controls are based is similarly restrictive. In particular, "depiction or discussion of debated social issues and related acts in negative or partisan context" – which would seemingly include much normal political debate – is deemed high-risk, and thus excluded by default settings (GARM, 2022).

This has serious implications for the freedom of public discourse, political activism and independent political commentary. Moreover, this not only affects independent creators, but also traditional media. Demonetisation of news content through brand safety tools is already widespread in display advertising, affecting the revenue of publishers covering serious news and topics relevant to marginalised groups, such as LGBTQIA+ issues (Iwańska, 2020; Check My Ads, 2020; Parker, 2021). Similarly, increased use of brand suitability tools which demonetise coverage of "debated social issues'' and "sensitive current events'' will likely affect its visibility on social media. Publishers' reliance on social media traffic to attract audiences can influence editorial decisions (Cornia et al., 2018; Petre, 2021). Consequently, this could not only financially undermine news journalism but also incentivise publishers to shift towards lighter, advertiser-friendly topics.

Cultural imperialism

These problematic dynamics are compounded by the global reach of social media. Dominant platforms exercise extensive power to regulate online speech around

the world (Kwet, 2019). Standardising brand suitability tools across these global platforms will effectively regulate ad placements worldwide in line with what largely Western corporations consider "appropriate". Leading platforms' corporate cultures and policies often reflect US cultural norms (Klonick, 2018; Monea, 2022). More importantly, however, US and Canadian consumers are the most valuable audiences for social media advertising, making it economically rational to prioritise attracting users and advertisers in these markets (Oversight Board, 2022, p. 31). Despite GARM's nominally global scope, its steering committee comprises four US-based and two UK-based multinational companies, and two US-based and one UK-based advertising trade associations (WFA, n.d.b). Accordingly, all major platform companies – including non-US-owned companies like TikTok – operate in a commercial and institutional environment that incentivises them to cater to the business needs and values of a US-dominated advertising industry.

This is particularly apparent in relation to sexual content policies, which largely reflect relatively conservative US mores, with little consideration for varying cultural norms around sex and nudity (Roberts, 2018; Are & Paasonen, 2021; Monea, 2022). For example, US-based platforms have been prominently criticised in several European countries for restricting depictions of nudity in artistic (Sotto, 2018; Hunt, 2021) or political contexts (Ritter et al., 2016). Moderation policies frame sex and especially sex work as risky or dangerous (Are & Briggs, 2023), and sex work-related content is often indiscriminately removed, even in countries where it is legal (Barwulor et al., 2021). Equally, it seems unlikely that brand suitability tools will reflect accurate or nuanced understandings of what is "sensitive", "political" or "partisan" in different national contexts – especially given major platforms' systematic underinvestment in moderation resources and AI capabilities for less-wealthy markets and languages other than English (Scheck et al., 2021).

Considering how brand suitability influences content curation more broadly, as well as the moderation of content deemed particularly risky, also raises questions about how this could shape online culture around the world in ways that reflect Western perspectives and interests. For example, the facial filters mentioned in section 2(d) typically promote ideals of beauty that are highly racialised, giving users lighter skin and more European-looking features (Griffin, 2023a). While generally concerning, this might have particularly negative effects in non-Western contexts, for example in countries where light-skinned beauty ideals have led to widespread use of harmful skin-lightening products (Jacobs et al., 2016).

Errors and bias

The above discussion suggests that brand safety and suitability tools will have several concerning policy implications even if they work exactly as intended. However, content classification tools are also inevitably prone to errors and bias (Chowdhury, 2022) – and given established patterns of algorithmic bias, it is likely that mistakes such as "false positive" classifications of content as unsafe will most heavily affect marginalised groups (Bishop, 2021). Brand safety tools used for text-dominated news sites are erratic and rife with racist and homophobic bias (CHEQ, 2019; Check My Ads, 2020). AI classifiers for the types of visual content that dominate social media are known to be even more limited and unreliable (Thakur & Llansó, 2021; Chowdhury, 2022).

In general, ad targeting tolerates high inaccuracy: the pools of potential content and audiences are so large that inadvertently excluding some false positives is unimportant (Wachter, 2020). Advertisers and platforms have little to lose from overinclusive brand safety measures. The side effect will be a systematic loss of resources and visibility for marginalised creators and non-mainstream or controversial perspectives.

The EU legal landscape

EU policymakers have taken action to address concerns around freedom of expression, discrimination and media freedom, like those discussed above – notably through the DSA, which explicitly regulates areas like content moderation, demonetisation and recommendations. This section thus evaluates how this legal framework relates to brand safety and suitability measures and could address the policy concerns described in section 4. It first assesses the DSA's generally-applicable provisions on content moderation, including demonetisation, before discussing some of the special obligations for large platforms, which could impact other brand safety measures. Finally, it discusses the 2022 Code of Practice on Disinformation, the only EU legal instrument explicitly mentioning (and encouraging) brand safety tools.

Content moderation

Articles 14-21 DSA establish various regulatory requirements for moderation systems.[3] Terms and conditions, including content policies, should be transparent and

3. Some provisions apply to all intermediary services, some only to hosting services, and some to the still-narrower category of "online platforms". As social media platforms fall into all three categories,

accessible (Article 14(1)). Where platforms restrict content – for example by deleting or demonetising it, or demoting it in recommendations – they must inform users of the decision and reasons (Article 17). Platforms must allow users to appeal such decisions through an "effective internal complaint-handling system", operated by "appropriately qualified staff, and not solely on the basis of automated means", and reverse decisions shown to be unfounded (Article 20). Unsatisfied users can further appeal to out-of-court dispute settlement bodies (Article 21).

Three general points can be made about these provisions' implications for brand safety measures. First, although individual procedural safeguards generally have limited capacities to address systemic issues like those discussed in section 4, professional creators' interests in brand safety may drive more widespread and strategic use of such procedures, amplifying their impact. Second, however, these provisions cannot address all the mechanisms identified in section 2. In particular, they do not appear to cover indirect effects of inventory controls on monetisability and visibility, and have limited application to recommendation systems generally. Finally, more broadly, their primary aim is ensuring fair and consistent application of platforms' policies. As they mostly do not substantively regulate the policies themselves, they do not fundamentally challenge platforms' incentives and abilities to regulate online communications and media in line with advertiser interests.

On the first point, individual procedural protections do not facilitate consideration and contestation of systemic issues and biases that lead to mistaken moderation decisions (Douek, 2022), or the organisational structures and policy objectives shaping moderation systems (Griffin, 2023b). They are also rarely used in practice (Urban et al., 2017), and since they demand time and digital literacy, are likely to primarily benefit more privileged users (Hoffmann, 2019).

However, in the context of brand safety, particularly demonetisation, these safeguards may have broader effects. Such policies are a major concern for professional creators, who have the motivation and expertise to effectively utilise procedural protections. In 2019, LGBTQIA+ YouTubers brought a class action lawsuit in California, alleging systematic discrimination in YouTube's enforcement of demonetisation and other policies (*Divino Group v Google* [2019]). They were unsuccessful, since US law gives platforms near-absolute discretion over content moderation, but this illustrates professional creators' capacity and willingness to utilise legal protections, including for collective goals – and the DSA will now give them more avenues to do so.

these differences in scope are unimportant for present purposes.

While Article 20's appeals process may not offer complete redress to individual creators (for example, it may be too late for a topical video to have its intended impact), large-scale uptake by professional creators of appeals procedures which are costly for platforms could incentivise general improvements to moderation systems, indirectly benefiting all users. Such pressures could also be amplified by some provisions whose scope goes beyond individual decisions. First, moderation policies must generally be clear and transparent. A demonetisation policy as vague as Meta's (n.d.d), which only states what "may affect monetisation", arguably violates this requirement. Under Article 53, creators could complain to national regulators to demand clearer guidance, which would in turn facilitate challenges to arbitrary decisions. Second, platforms must publish regular transparency reports on their moderation systems, including the use of automation (Article 15). In combination with Article 40(4), which requires the largest platforms to provide internal data to researchers, and Article 27, which requires transparency about the 'main parameters' of recommendation systems, this could enable further scrutiny and public criticism of brand safety measures by creators, researchers and other stakeholders.

Finally, Article 14(2) requires moderation policies to be applied "with due regard to" fundamental rights, including non-discrimination, freedom of expression, and media freedom and pluralism. What exactly it means for platforms to have regard to these abstract principles remains highly uncertain (Griffin, 2023b). However, this could create additional avenues for creators and other engaged stakeholders to challenge brand safety policies (Quintais et al., 2022), for example by complaining to regulators under Article 53 that anti-LGBTQIA+ discrimination violates Article 14.

Nonetheless, these provisions' scope is limited in key respects. In particular, while demonetisation is subject to procedural protections, inventory exclusions in principle follow from advertiser choice, not platform policies, so might not be covered – even though exclusion from standard inventory tiers might have similar effects to complete demonetisation. Recital 55 also states that "monetisation…can be restricted by suspending or terminating the monetary payment or revenue associated to that information". Thus, the DSA appears to define demonetisation narrowly, limited to contexts where it affects revenue-sharing schemes. However, as section 2(b) noted, even where revenue is not shared, completely or partially excluding content from ad placements may still affect visibility and therefore raises policy concerns.

While Articles 17 and 20 apply to "restrictions of visibility", their scope is again

somewhat unclear. Article 3(t) suggests that visibility restrictions only qualify as moderation when they are aimed at enforcing laws or policies, not when recommendation systems are generally optimised to pursue commercial objectives like maximising ad revenue. Identifying policy-based demotions, which generally operate as a discrete stage of recommendation processes, is also more technically feasible (Leerssen, 2023). However, (de)monetisation will likely be one of many data points factored into rankings. How much must it decrease the overall ranking to qualify as a demotion? Even if that question had a correct answer in principle, how could that be assessed in practice, given the complexity and opacity of recommendation algorithms? In any case, the effects of such practices would principally unfold probabilistically and at scale – the problem not being that individual posts do not achieve their rightful visibility, but rather that certain content types are generally less likely to become widely visible, and thus also less likely to be produced by creators. The DSA's individualistic procedural safeguards cannot effectively address these dynamics (Griffin, 2023b).

This points to a broader limitation. These provisions provide avenues to challenge erroneous, arbitrary and discriminatory application of moderation and demonetisation policies, but mostly do not address the substantive policies themselves. Platforms can still ban unsafe content; follow GARM's restrictive suitability definitions; and prioritise recommending the most easily-commodifiable content – as long as they do so in accordance with clear and consistent policies.

The key exception to this is Article 14(2)'s requirement to have regard to fundamental rights. However, this requirement is so vague its effects may be minimal. Essentially all moderation policies and decisions involve multiple competing rights, including platforms' own freedom to run a business; requiring platforms to explain how they considered and balanced various rights will not be a meaningful substantive constraint (Griffin, 2023b). Much will depend on how prescriptively regulators interpret platforms' fundamental rights obligations (Quintais et al., 2022). Overall, however, Articles 14-21 DSA appear inadequate to address the systemic problems identified in this article.

Very large online platforms

The DSA also attempts to address systemic issues – in particular, through extended obligations for "very large online platforms", with over 45 million EU users (Article 33(1)). Article 34 requires them to regularly assess "systemic risks" to various public-interest objectives, including fundamental rights (Article 34(1)(b)). They must consider factors like recommendations and "systems for selecting and pre-

senting advertisements" (Article 34(2)). The scope is therefore much broader than moderation, extending to inventory controls, recommendation algorithms and other design features. Platforms must take proactive measures to mitigate identified risks, for example by redesigning relevant algorithmic systems (Article 35); these measures are subject to yearly independent audits (Article 37) and oversight by the Commission (Section 4). In principle, discrimination, suppression of political debate, cultural imperialism and bias seem exactly the kinds of systemic risks to media freedom, non-discrimination, free speech and other fundamental rights that these provisions aim to address.

That said, it remains highly unclear how they will be enforced and what concrete effects they will have. Where companies are responsible for managing and documenting compliance, through mechanisms like risk assessments and audits, they have extensive freedom to define and deal with risks in self-serving ways – typically constructing risks as threats arising when things go wrong, rather than social harms flowing from "normal" business practices (Cohen, 2019; Waldman, 2020). In this context, platforms' risk assessments will likely focus on discrete problems which are bad for business, like terrorist content, rather than systemic harms resulting from their business models. This may be compounded by the novelty of the relevant provisions, which create vaguely-defined obligations in unfamiliar areas. Given the lack of clear, widely-accepted standards on compliance, risk mitigation and auditing methodologies, this oversight system may be susceptible to corporate capture (Laux et al., 2021).

As with Article 14(2), much will depend on the Commission's policies and objectives, as the primary regulator of large platforms. Risk assessment and mitigation obligations could be little more than formalities; they could also present opportunities to demand significant, substantive policy and design changes addressing issues like media pluralism, anti-LGBTQ+ discrimination and freedom of political debate. However, as the following subsection suggests, the Commission may not be inclined to take such an approach.

The Code of Practice on Disinformation

In 2022, the Commission and leading platform companies, adtech businesses and ad industry associations (including GARM and the WFA) agreed to an expanded version of the 2018 Code of Practice on Disinformation (CoP). The CoP is non-binding, and the previous iteration was not considered particularly effective (Sander, 2021). However, the new CoP is significantly more detailed, and will become an official code of conduct under Article 45 DSA; it will therefore have more teeth, as

regulators will use it in assessing compliance with legal obligations, in particular very large platforms' risk mitigation obligations.

Notably, the new CoP addresses brand safety tools in some detail. It frames them as an important lever to prevent monetisation of disinformation – one that already works well, but must be further expanded: "Avoiding the misplacement of advertising on online disinformation websites requires further refinement of already widely used brand safety tools to successfully continue to meet this challenge" (Section II.h). In Section II, platforms commit to improve demonetisation and ad-tracking tools; advertisers commit to use brand safety tools in media planning, buying and reporting, and preferentially use ad vendors with effective brand safety measures; and brand safety software providers commit to "outline how they are ensuring transparency and appealability about their processes and outcomes".

Insofar as the CoP successfully encourages further development and uptake of brand safety measures, this will have consequences beyond the disinformation context. Where platforms develop new moderation tools for specific areas, it will generally make technical and commercial sense to roll them out across moderation systems (Elkin-Koren & Perel, 2020). Advertisers will likely demand that expanded controls and tracking tools should also address other problematic content categories – something platforms are also incentivised to do, since they can present this as a further mitigation measure for other systemic risks mentioned in Article 34.

The CoP also entrenches the market logic in which platforms' and creators' dependence on advertising revenue empowers advertisers as "censors" (Baker, 1992, p. 2009) – effectively legitimising this censorship power and charging them with using it responsibly. The cursory mention of transparency and appealability also seems aimed at strengthening perceived legitimacy. The CoP suggests that the Commission envisages a future where the flow of advertising money continues to be a primary structural force shaping social media governance. It may therefore be unlikely to push for substantial changes to brand safety practices.

Conclusion

Advertisers' brand safety concerns are an underappreciated force shaping social media governance, with worrying implications for media freedom and diversity. Advertiser influence seems to be increasing, encouraging the rollout of standardised norms and tools across major platforms. Such trends do not just affect individual creators: they have much broader implications for online media, as monetis-

ability and advertiser appeal become key factors shaping what becomes widely visible.

As such, the terminological shift to brand suitability suggests much more than better, customisable controls for advertisers. It evokes a broader logic of curation that appears increasingly dominant, where all social media content is organised based on what major brands consider suitable for their commercial objectives. Ultimately, this will tend to reinforce the gender, race and class inequalities that have always shaped access to the media (Bishop, 2018, 2021).

The DSA provides some promising new tools to address arbitrary and unfair content removal and demonetisation. However, it does not adequately address these industry-wide structural forces shaping how platforms structure, curate and distribute content. Similarly, while the literature on social media governance has engaged with cross-industry trends like standardisation of content policies (Douek, 2020), civil society involvement (Dvoskin, 2022) and the influence of law enforcement (Bloch-Wehba, 2022), there has been insufficient engagement with the structural incentives and dynamics of the advertiser-funded industry. The concept of brand suitability thus offers a useful lens for deeper scholarly engagement with the political economy of social media.

ACKNOWLEDGEMENTS

I would like to thank Annina Claesson, Daphne Keller and the organisers and participants of the 2023 PlatGov Research Network Conference for helpful discussions and comments.

References

Alexander, J. (2019, September 30). YouTube moderation bots punish videos tagged as 'gay' or 'lesbian,' study finds. *The Verge*. https://www.theverge.com/2019/9/30/20887614/youtube-moderation-lgbtq-demonetization-terms-words-nerd-city-investigation

Are, C., & Briggs, P. (2023). The emotional and financial impact of de-platforming on creators at the margins. *Social Media + Society*, *9*(1), 205630512311551. https://doi.org/10.1177/20563051231155103

Are, C., & Paasonen, S. (2021). Sex in the shadows of celebrity. *Porn Studies*, *8*(4), 411–419. https://doi.org/10.1080/23268743.2021.1974311

Baker, C. E. (1992). Advertising and a democratic press. *University of Pennsylvania Law Review*,

6(140), 2097–2243. https://doi.org/10.2307/3312414

Barwulor, C., McDonald, A., Hargittai, E., & Redmiles, E. M. (2021). "Disadvantaged in the American-dominated internet": Sex, work, and technology. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445378

Bayley, C. (2021, May 25). Sexual censorship on social media: What I learned [Belfer Center for Science and International Affairs]. *Perspectives on Public Purpose: For Emerging Technologies*. https://www.belfercenter.org/publication/sexual-censorship-social-media-what-i-learned

Bishop, S. (2018). Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm. *Convergence*, *24*(1), 69–84. https://doi.org/10.1177%2F1354856517736978

Bishop, S. (2021). Influencer management tools: Algorithmic cultures, brand safety, and bias. *Social Media + Society*, *7*(1), 205630512110030. https://doi.org/10.1177/20563051211003066

Bloch-Wehba, H. (2022). Content moderation as surveillance. *Berkeley Journal of Law & Technology*, *36*(3), 1297–1340. https://doi.org/10.15779/Z389C6S202

Caplan, R., & Gillespie, T. (2020). Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media + Society*, *6*(2), 1–13. https://doi.org/10.1177/2056305120936636

Carah, N., & Brodmerkel, S. (2022). Regulating platforms' algorithmic brand culture: The instructive case of alcohol marketers on social media. In T. Flew & F. R. Martin (Eds.), *Digital Platform Regulation* (pp. 111–130). Springer International Publishing. https://doi.org/10.1007/978-3-030-95220-4_6

Check My Ads. (2020). *Inside the chaos of brand safety technology* (Branded) [Newsletter]. https://checkmyads.org/branded/inside-the-chaos-of-brand-safety/

CHEQ. (2019). *Brand safety's technological challenge: How keyword blacklists are killing reach and monetization. A study by CHEQ's Department of Data Science* [Study]. CHEQ. https://info.cheq.ai/hubfs/Research/Brand_Safety_Blocklist_Report.pdf

Chowdhury, N. (2022). *Automated content moderation: A primer* (Program on Platform Regulation) [Primer]. Stanford Cyber Policy Center. https://cyber.fsi.stanford.edu/news/automated-content-moderation-primer

Cohen, J. E. (2019). *Between truth and power: The legal constructions of informational capitalism* (1st ed.). Oxford University Press. https://doi.org/10.1093/oso/9780190246693.001.0001

Cornia, A., Sehl, A., Levy, A., & Nielsen, R. K. (2018). *Private sector news, social media distribution, and algorithm change* (Digital News Project) [Report]. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/our-research/private-sector-news-social-media-distribution-and-algorithm-change

Cotter, K. (2019). Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society*, *21*(4), 895–913. https://doi.org/10.1177/1461444818815684

Creator Insider. (2018, December 9). *YouTube algorithm questions explained by YouTube employees (part 3)* [Video]. YouTube. https://www.youtube.com/watch?v=fxRIUFyv_Rk&ab_channel=CreatorInsider

Cunningham, S., & Craig, D. (2019). Creator governance in social media entertainment. *Social Media*

+ *Society*, *5*(4), 205630511988342. https://doi.org/10.1177/2056305119883428

DoubleVerify. (2021, May 25). Brand safety vs. Brand suitability. *DV Publisher Insights*. https://pub.doubleverify.com/blog/brand-safety-vs-brand-suitability/

Douek, E. (2020). The rise of content cartels. Knight first amendment institute: The tech giants, monopoly power, and public discourse. *Knight First Amendment Institute Essays & Scholarship*. https://perma.cc/H6HZ-NWS7

Douek, E. (2022). Content moderation as systems thinking. *Harvard Law Review*, *136*(2), 526–607. https://doi.org/10.2139/ssrn.4005326

Dvoskin, B. (2022). Representation without elections: Civil society participation as a remedy for the democratic deficits of online speech governance. *Villanova Law Review*, *67*(3), 447–508. https://doi.org/10.2139/ssrn.3986181

Elkin-Koren, N., & Perel, M. (2020). Separation of functions for AI: Restraining speech regulation by online platforms. *Lewis & Clark Law Review*, *24*(3), 857–898. https://doi.org/10.2139/ssrn.3439261

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press. https://doi.org/10.12987/9780300235029

Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media + Society*, *8*(3), 205630512211175. https://doi.org/10.1177/20563051221117552

Glatt, Z. (2022). 'We're all told not to put our eggs in one basket': Uncertainty, precarity and cross-platform labor in the online video influencer industry. *International Journal of Communication*, *16*, 3853–3871. https://ijoc.org/index.php/ijoc/article/view/15761

Glatt, Z., & Banet-Weiser, S. (2021). Productive ambivalence, economies of visibility, and the political potential of feminist YouTubers. In S. Cunningham & D. Craig (Eds.), *Creator culture: An introduction to global social media entertainment* (pp. 39–56). New York University Press. https://doi.org/10.18574/nyu/9781479890118.003.0006

Global Alliance for Responsible Media. (2022). *GARM brand safety floor + suitability framework* [White paper]. World Federation of Advertisers. https://wfanet.org/knowledge/item/2022/06/17/GARM-Brand-Safety-Floor--Suitability-Framework-3

Goanta, C. (2023). The new social media: Contracts, consumers and chaos. *Iowa Law Review*, *108*, 118–130.

Google. (n.d.). *About content exclusions for video campaigns*. Google Ads Help. https://support.google.com/google-ads/answer/7515513?hl=en

Griffin, R. (2023, March 24). TikTok's confidence-destroying bold glamour filter is the logical product of platforms built for consumerism. *Tech Policy Press*. https://techpolicy.press/tiktoks-confidence-destroying-bold-glamour-filter-is-the-logical-product-of-platforms-built-for-consumerism/

Griffin, R. (2023b). Rethinking rights in social media governance: Human rights, ideology and inequality. *European Law Open*, *2*(1), 30–56. https://doi.org/10.1017/elo.2023.7

Hamilton, J. T., & Morgan, F. (2018). Poor information: How economics affects the information lives of low-income individuals. *International Journal of Communication*, *12*, 2832–2850. https://ijoc.org/index.php/ijoc/article/view/8340/2399

Hendrix, J. (2021, May 19). Researchers detail Sleeping Giants strategy against misinformation. *Tech*

Policy Press. https://techpolicy.press/researchers-detail-sleeping-giants-strategy-against-misinformation/

Herman, E. S., & Chomsky, N. (1995). *Manufacturing consent: The political economy of the mass media*. Vintage Books.

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Hunt, E. (2021, October 16). Vienna museums open adult-only OnlyFans account to display nudes. *The Guardian*. https://www.theguardian.com/artanddesign/2021/oct/16/vienna-museums-open-adult-only-onlyfans-account-to-display-nudes

Iwańska, K. (2020). *To track or not to track? Towards privacy-friendly and sustainable online advertising* [Research brief]. Panoptykon Foundation. https://panoptykon.org/sites/default/files/publikacje/panoptykon_to_track_or_not_to_track_final.pdf

Jacobs, M., Levine, S., Abney, K., & Davids, L. (2016). Fifty shades of African lightness: A bio-psychosocial review of the global phenomenon of skin lightening practices. *Journal of Public Health in Africa*, *7*(2), 67–70. https://doi.org/10.4081/jphia.2016.552

JoergSprave. (2018, September 21). *Debunked: YouTube caught lying! (YouTubers Union video)* [Video]. YouTube. https://www.youtube.com/watch?v=Tn5rOOfW7bc&ab_channel=JoergSprave

Klonick, K. (2018). The new governors: The people, rules, and processes governing online sdpeech. *Harvard Law Review*, *131*(6), 1598–1670.

Kumar, S. (2019). The algorithmic dance: YouTube's Adpocalypse and the gatekeeping of cultural content on digital platforms. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1417

Kwet, M. (2019). Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class*, *60*(4), 3–26. https://doi.org/10.1177/0306396818823172

Laux, J., Wachter, S., & Mittelstadt, B. (2021). Taming the few: Platform regulation, independent audits, and the risks of capture created by the DMA and DSA. *Computer Law & Security Review*, *43*, 105613. https://doi.org/10.1016/j.clsr.2021.105613

Lee, C., Kim, J., & Lim, J. S. (2021). Spillover effects of brand safety violations in social media. *Journal of Current Issues & Research in Advertising*, *42*(4), 354–371. https://doi.org/10.1080/10641734.2021.1905572

Leerssen, P. (2023). An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. *Computer Law & Security Review*, *48*, 1–13. https://doi.org/10.1016/j.clsr.2023.105790

Madio, L., & Quinn, M. (2023). *Content moderation and advertising in social media platforms*. SSRN. https://doi.org/10.2139/ssrn.3551103

Marvin, G., & Meisel, S. (2017). Protecting the brand in the era of fake news: Why brands need advertisement verification tools. *Journal of Digital & Social Media Marketing*, *5*(4), 322–331.

Meehan, E. (2006). Gendering the commodity audience: Critical media research, feminism and political economy. In M. G. Durham & D. M. Kellner (Eds.), *Media and Cultural Studies KeyWorks Revised Edition* (2nd ed., pp. 311–321). Blackwell Publishing.

Meta. (2019, April 10). *Updating our brand safety controls*. Facebook Business. https://www.facebook.com/business/news/updating-our-brand-safety-controls

Meta. (2022, March 17). *Brand suitability updates: Third-party verification and content-based controls for feed*. Meta Announcements. https://www.facebook.com/business/news/brand_suitability_updates

Meta. (n.d.a). *Brand safety controls*. Meta Business Help Centre. https://www.facebook.com/business/help/1926878614264962?id=1769156093197771

Meta. (n.d.b). *About brand safety on Facebook, Instagram, WhatsApp and audience network*. Meta Business Help Centre. https://www.facebook.com/business/help/1559334364175848?id=1769156093197771

Meta. (n.d.c). *Facebook brand safety description of methodology*. Meta Business Help Centre. https://www.facebook.com/business/help/4360253257396424?id=1769156093197771

Meta. (n.d.d). *Content monetisation policies*. Meta Business Help Centre. https://www.facebook.com/business/help/1348682518563619?id=2520940424820218

Meta. (n.d.e). *About inventory filter*. Meta Business Help Centre. https://www.facebook.com/business/help/3001448133206080?id=1769156093197771

Milmo, D. (2023, February 20). Does paid-for Facebook and Instagram signal end of free-access orthodoxy? *The Guardian*. https://www.theguardian.com/technology/2023/feb/20/facebook-instagram-paid-for-signal-free-access-meta

Monea, A. (2022). *The digital closet: How the internet became straight*. MIT Press. https://doi.org/10.7551/mitpress/12551.001.0001

Morra, J. (2021, September 22). GARM brand suitability: Developing a collaborative interpretation of the standard. *IAB Tech Lab Blog*. https://iabtechlab.com/blog/garm-brand-suitability-developing-a-collaborative-interpretation-of-the-standard/

Narayanan, A. (2023). Understanding social media recommendation algorithms. *Knight First Amendment Institute Essays & Scholarship*. https://perma.cc/F3NP-FEQX

Oasis Consortium. (n.d.). *Brand safety vs. Brand suitability: What's the difference?* [White paper]. https://www.oasisconsortium.com/insights/brand-safety-vs-brand-suitability

Oversight Board. (2022). *Oversight board publishes policy advisory opinion on Meta's cross-check programme* [Policy review summary]. https://www.oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s-cross-check-program/

Parker, B. (2021, January 27). How advertisers defund crisis journalism. *The New Humanitarian*. https://www.thenewhumanitarian.org/analysis/2021/01/27/brand-safety-ad-tech-crisis-news

Petre, C. (2021). *All the news that's fit to click: How metrics are transforming the work of journalists*. Princeton University Press. https://doi.org/10.1515/9780691228754

Quintais, J. P., Appelman, N., & Fahy, R. (2022). Using terms and conditions to apply fundamental rights to content moderation. *German Law Journal*, Advance online publication. https://doi.org/10.2139/ssrn.4286147

Riemer, K., & Peter, S. (2021). Algorithmic audiencing: Why we need to rethink free speech on social media. *Journal of Information Technology*, *36*(4), 409–426. https://doi.org/10.1177/02683962211013358

Ritter, K., Ortutay, B., & Anderson, M. (2016, September 9). Facebook allows postings of 'napalm girl' photo after debate. *AP News*. https://apnews.com/article/technology-europe-asia-pacific-nick-ut-denmark-3cab71acfabc4a228db9ced888ea4383

Roberts, S. T. (2018). Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday*, *23*(3). https://doi.org/10.5210/fm.v23i3.8283

Rodriguez, J. A. (2022). LGBTQ incorporated: YouTube and the management of diversity. *Journal of Homosexuality*, *70*(9), 1807–1828. https://doi.org/10.1080/00918369.2022.2042664

Sander, B. (2021). Democratic disruption in the age of social media: Between marketized and structural conceptions of human rights law. *European Journal of International Law*, *32*(1), 159–193. https://doi.org/10.1093/ejil/chab022

Scheck, J., Purnell, N., & Horwitz, J. (2021, September 16). Facebook employees flag drug cartels and human traffickers. The company's response is weak, documents show. *Wall Street Journal*. https://www.wsj.com/amp/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953

Seufert, E. B. (2023, January 11). The app tracking transparency recession. *Mobile Dev Memo*. https://mobiledevmemo.com/the-att-recession/

Shepherd, T. (2014). Gendering the commodity audience in social media. In C. Carter, L. Steiner, & L. McLaughlin (Eds.), *The Routledge Companion to Media and Gender* (1st ed., pp. 157–167). Routledge. https://doi.org/10.4324/9780203066911

Sloane, G. (2022). *Meta Picks Zefr as first news feed brand safety measurement partner* [Press release]. https://zefr.com/press/meta-picks-zefr-as-first-news-feed-brand-safety-measurement-partner

Smythe, D. (2006). On the audience commodity and its work. In M. G. Durham & D. M. Kellner (Eds.), *Media and Cultural Studies KeyWorks Revised Edition* (2nd ed., pp. 230–256). Blackwell Publishing.

Sotto, P. (2018, March 16). French court issues mixed ruling in Facebook nudity case. *AP News*. https://apnews.com/article/ebbd9a846504460ea184201dccce303d

Southerton, C., Marshall, D., Aggleton, P., Rasmussen, M. L., & Cover, R. (2020). Restricted modes: Social media, content classification and LGBTQ sexual citizenship. *New Media & Society*, *23*(5), 920–938. https://doi.org/10.1177%2F1461444820904362

Steinem, G. (1990). Sex, Lies & Advertising. *Ms*, Article 28. https://www1.udel.edu/comm245/readings/advertising.pdf

Thakur, D., & Llansó, E. (2021). *Do you see what I see? Capabilities and limits of automated multimedia content analysis* [Research report]. Center for Democracy & Technology. https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/

Thomasen, K. (2023). Safety in artificial intelligence & robotics governance in Canada. *Canadian Bar Review*, Advance online publication. https://doi.org/10.2139/ssrn.4387804

Tiidenberg, K. (2021). Sex, power and platform governance. *Porn Studies*, *8*(4), 381–393. https://doi.org/10.1080/23268743.2021.1974312

TikTok. (2022, May 18). *Introducing TikTok branded mission: Inspiring brand and creator collaborations*. TikTok. https://newsroom.tiktok.com/en-us/introducing-tiktok-branded-mission-inspiring-brand-and-creator-collaborations

TikTok. (n.d.). *TikTok inventory filters*. TikTok Business Help Center. https://ads.tiktok.com/help/article?aid=10011439

Urban, J. M., Schofield, B. L., & Karaganis, J. (2017). Takedown in two worlds: An empirical analysis. *Journal of the Copyright Society of the USA*, *64*, 483–520. https://doi.org/10.31235/osf.io/mduyn

van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199970773.001.0001

Wachter, S. (2020). Affinity profiling and discrimination by association in online behavioral advertising. *Berkeley Technology Law Journal*, *35*(2), 367–430. https://doi.org/10.15779/Z38JS9H82M

Waldman, A. E. (2020). Privacy law's false promise. *Washington University Law Review*, *97*(3), 773–834. https://doi.org/10.2139/SSRN.3339372

Waldman, A. E. (2021). Disorderly content. *Washington Law Review*, *97*. https://doi.org/10.2139/ssrn.3906001

Wodinsky, S. (2020, June 26). The 'stop hate for profit' movement isn't going to stop anything. *Gizmodo*. https://gizmodo.com/the-stop-hate-for-profit-movement-isnt-going-to-stop-1844147197

Wolinsky, D. (2022, May 12). YouTube receives brand safety distinction for second year. *Google Ads & Commerce Blog*. https://blog.google/products/ads-commerce/youtube-brand-safety-second-accreditation/

World Federation of Advertisers. (2020). *WFA and platforms make major progress to address harmful content* [Press release]. https://wfanet.org/knowledge/item/2020/09/23/WFA-and-platforms-make-major-progress-to-address-harmful-content

World Federation of Advertisers. (n.d.a). *Global alliance for responsible media*. https://wfanet.org/leadership/garm/about-garm

World Federation of Advertisers. (n.d.b). *GARM Members*. https://wfanet.org/leadership/garm/members-governance

Zefr. (n.d.a). *Zefr for TikTok*. Zefr. https://zefr.com/product/zefr-for-tiktok

Zefr. (n.d.b). *Zefr for Facebook in-stream video*. Zefr. https://zefr.com/product/zefr-for-facebook

Zefr. (n.d.c). *Brand suitability*. Zefr. https://zefr.com/brand-suitability

Zefr. (n.d.d). *Zefr's data solutions*. Zefr. https://zefr.com/product

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, *14*(1), 79–95. https://doi.org/10.1002/poi3.287

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

cnrs centre internet et societe

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies