INTERNET POLICY REVIEW
Journal on internet regulation

Volume 12 **Issue** 1

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Humour as an online safety issue: Exploring solutions to help platforms better address this form of expression

Ariadna Matamoros-Fernández *Queensland University of Technology*
Louisa Bartolo *Queensland University of Technology*
Luke Troynar *Queensland University of Technology*

DOI: https://doi.org/10.14763/2023.1.1677

Published: 25 January 2023
Received: 7 September 2021 **Accepted:** 7 March 2022

Funding: Ariadna Matamoros-Fernández, together with Amy Johnson, received funding from Facebook in a competitive grant scheme for the research underpinning this article.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.
Licence: This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 License (Germany) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. https://creativecommons.org/licenses/by/3.0/de/deed.en
Copyright remains with the author(s).

Citation: Matamoros-Fernández, A., Bartolo, L., & Troynar, L. (2023). Humour as an online safety issue: Exploring solutions to help platforms better address this form of expression. *Internet Policy Review*, *12*(1). https://doi.org/10.14763/2023.1.1677

Keywords: Humour, Online safety, Social media practices, Content moderation

Abstract: This paper makes a case for addressing humour as an online safety issue so that social media platforms can include it in their risk assessments and harm mitigation strategies. We take the 'online safety' regulation debate, especially as it is taking place in the UK and the European Union, as an opportunity to reconsider how and when humour targeted at historically marginalised groups can cause harm. Drawing on sociolegal literature, we argue that in their online safety efforts, platforms should address lawful humour targeted at historically marginalised groups because it can cause individual harm via its cumulative effects and contribute to broader social harms. We also demonstrate how principles and concepts from critical humour studies and Feminist Standpoint Theory can help platforms assess the differential impacts of humour.

Introduction

Humour, broadly understood as the act of making fun of seriousness, or an utterance that aims to cause amusement (Lockyer & Pickering, 2005), is a key element of social media culture. People often participate online to have a laugh (Shifman, 2014), and communities mobilise humorous expression to build solidarity and speak truth to power (Brown, 2019; Carlson & Frazer, 2021). But humour can also be used to silence and discriminate and, on social media, historically marginalised groups are frequently targets (Greene, 2019; Phillips & Milner, 2017). The policies and content moderation processes of digital platforms are currently not well equipped to recognise the harms derived from humour, nor to distinguish it from other forms of expression – decisions that are always relative to specific cultural contexts and complex due to humour's inherently ambiguous nature (Kuipers, 2011; Weaver, 2011). There is the risk, therefore, that content moderation practices may result in the removal of important critical or harmless humour (Dias Oliva et al., 2020; Paasonen et al., 2019) and/or fail to effectively moderate humour that sows division and hate (Fielitz & Moore, 2018; Matamoros-Fernández, 2017). The challenge of moderating humour is exacerbated by the reluctance of tech companies to limit the flow of highly engaging and profitable controversial humorous content, such as viral videos trading in racist stereotypes (Roberts, 2016). Additionally, while platforms tend to pay insufficient attention to power differentials among protected categories in their governance of speech and conduct (e.g., hate speech policies treat 'race' as a single, undifferentiated category) (Bartolo, 2021; Siapera & Viejo-Otero, 2021), we argue it is important to attend to contextual structural factors when assessing the harms associated with content/behaviour which may not, by itself, rise to the level of illegal hate speech or crime, including certain forms of humorous expression.

In this paper, we take the 'online safety' regulation debate, especially as it is taking place in the UK and the European Union, as an opportunity to address humour as an online safety issue so that digital platforms can include it in their risk assessments and harm mitigation strategies. At a time when key advocacy groups are urging digital platforms to "engage with the particularly complex issues in online hate, such as self-hatred; truth and validity; and *humour and irony*" (Vidgen et al., 2021, our emphasis), it is critical to address humorous expression in relation to online safety and well-being. In particular, we argue that proposals that seek to regulate 'legal but harmful' content on digital platforms open up opportunities to better address humour that harms historically marginalised groups. 'Legal but harmful'[1] content/behaviour is a category originally included in the UK's Draft Online

Safety Bill, which was meant to push platforms to address content that is not illegal but is thought to be harmful in certain instances. In an amended version of the Bill[2], the UK government removed the 'legal but harmful' provisions and replaced them "with new duties to boost free speech and increase accountability of tech firms" (UK's Department for Digital, Culture, Media & Sport, 2022). The UK Online safety Bill draft and the intense debates it has spurred around the inclusion of 'legal but harmful' provisions raise critical issues far beyond the UK around how to best protect users via emerging online safety regulation. As we have argued elsewhere, "these debates encourage close examination of whether legal 'harm' frameworks are always the most appropriate for debating all online harms, and if not, a reflection on when legal frameworks might reach their limits." (Bartolo & Matamoros-Fernández, forthcoming). While the inclusion of 'legal but harmful' content within previous drafts of the UK Online Safety Bill received extensive backlash because of its implications for free speech and fears over state censorship of legitimate expression (see, for example, Digital, Culture, Media and Sport Committee, 2022, p. 11), it is in fact reflective of a pre-existing and growing trend to view platforms as not only capable of, but responsible *for*, dealing with harms in more expansive terms than legal frameworks currently support. In various cases, platforms' policies already go beyond what is expected of them by law (e.g., with their rules around misinformation and coordinated manipulation campaigns) (Gillespie, 2018; Suzor, 2019). The extensive community rules of some platforms, especially Facebook, are often the result of the continuous work of activist groups that have pushed tech companies to update, clarify and better enforce their community guidelines and decision-making procedures, especially in relation to how content affects historically marginalised groups (e.g. Facebook Safety, 2013; Ghaffary, 2020). The concern has been that these policies have often been introduced in ad hoc ways at platforms' whims, mostly in response to highly visible controversies and outside pressure; what Ananny and Gillespie (2016) call 'public shocks'. Moreover, the quality and consistency of enforcement has been impossible to evaluate systematically from the outside, and the limited 'Transparency Reports' produced by platforms have not satisfied growing calls for *real* 'Big Tech' accountability (Suzor et al., 2019). In any case, it is precisely because platforms already regulate con-

1. By this we mean the parts of the Bill that referred to 'content that is harmful to adults', as opposed to those related to 'illegal content' or 'content that is harmful to children'.

2. At the time of writing, a new draft of the UK Online Safety Bill was reintroduced and discussed in the UK Parliament in January 2023. In this amended version, for content categories that do not meet criminal thresholds–"such as the glorification of eating disorders, racism, anti-semitism or misogyny,"– the UK's government wants internet companies to offer more user controls to help people avoid seeing this type of abuse. In earlier versions, the Bill included provisions to force platforms to carry out risk assessments to document and address the risks of harm occurrence online for 'legal but harmful' content and behaviour.

tent (albeit neither consistently nor transparently) beyond what is legally required of them that some new online safety proposals (e.g., the UK's) have attempted – with little success – to create clearer guidelines around how platforms conceptualise and address 'legal but harmful' content and behaviour.

Somewhat similarly, the EU Digital Services Act (DSA) proposes obligations for 'Very Large Online Platforms/VLOPs' to assess and mitigate the risks associated with "harmful content (*which might not be illegal*) and the spread of disinformation" (European Parliament, 2022, emphasis our own). The EU identifies large platforms (those that reach 10% of the EU population, such as Meta, Twitter and Google) as carrying "systemic" and "societal" risks such as the undermining of fundamental rights (e.g., freedom of expression, non-discrimination, privacy) (DSA, pp. 31-32). As such, the DSA proposes transparency, risk assessment and independent audit requirements for large platforms, pushing them to identify and adopt proportionate measures to mitigate foreseeable risks of harm. In its current form, examples of lawful harms that the DSA cites relate to health, security, democratic elections, and disinformation. We believe that this opens up opportunities to specifically consider 'legal but harmful' humorous expression targeted at historically marginalised individuals and groups—its societal risks and its impact on people's safety.

In this paper, we acknowledge the complexities of identifying and assessing online humorous expression in general, yet want to specifically make the case for the importance of addressing 'legal but harmful' humour targeted at historically marginalised groups within social media platforms' online safety efforts. We argue that lawful humour targeted at historically marginalised groups can cause individual harm via its cumulative effects, and also contribute to broader social harms (e.g., by undermining social and political equality). We use the term 'historically marginalised groups', rather than 'protected groups', based on categories such as race, gender and sexuality, as is included in most anti-discrimination legislation. This is because we are interested in how new regulatory proposals for online safety can nudge[3] platforms to protect those individuals and groups that have historically

3. We acknowledge that governments and the state have serious potential to perpetrate harm themselves, and so relying on them to regulate platforms runs the risk of exacerbating certain forms of harm, including the harms of systemic inequality which this paper is particularly concerned with. In addition, even with the best intentions, governments are likely to find it exceedingly difficult to anticipate the broad range of harms that may arise across platforms (Bunting, 2018). This is why procedural accountability frameworks (as we are seeing in the EU in the form of risk assessments and transparency requirements for platforms) may be more appropriate for incentivising responsible platform governance whilst minimising overreach by both national regulators and platforms (Bunting, 2018). Importantly, however, as others have suggested (e.g. Schoenebeck & Blackwell, 2021), procedural accountability frameworks would benefit from being designed in collaboration with those most targeted by online harms.

been, and still are, subjected to systemic discrimination, which may include groups that are not (yet) recognised as such by particular countries' laws (McGowan, 2009).[4] To be sure, humour that does not target historically marginalised groups might also legitimately be considered 'legal but harmful' content in certain instances: for example, self-harm or suicide jokes, or jokes that have the potential to misinform. This kind of humour deserves the attention of regulators and platforms. However, in this paper we limit our focus to 'legal but harmful' humour *targeted at historically marginalised groups*.

We define this subcategory of 'legal but harmful' humour as any humorous communication that targets historically marginalised individuals and groups in a way that undermines their assurance as to equal status in the community. In this definition, and as we further unpack in this article, considering the 'speaker's power' in humorous interactions is relevant. Since we address 'legal but harmful humour' within the context of digital platforms, by 'speaker' we mean both the creators of humorous content and those who share that content. Lawful humour targeted at historically marginalised groups can be harmful regardless of speaker intent and irrespective of whether or not the direct target subjectively experiences upset. While the subjective experience of targets does matter, we argue it is not necessary as evidence of (risk of) harm, since there are broader social harms attached to humour that undermine historically marginalised groups' assurance as to equal status in the community (e.g., undermining values of equality).

We take inspiration from Australian Political Science scholar Katharine Gelber's (2019) systemic discrimination approach to defining 'hate speech' in our conceptualisation of 'legal but harmful' humour targeted at historically marginalised groups. Gelber (2019) argues that contexts of systemic inequality and the speaker's 'authority' matter significantly in determining speech's potential to harm. In fact, Gelber (2019, p. 7) challenges classic liberal conceptualisations of harm (e.g. Feinberg, 1987 – as we will discuss later in this paper) for not engaging in how certain speakers have greater capacity to harm than others. Whilst we acknowledge that Gelber's definition of 'hate speech' would include some jokes targeted at systematically marginalised groups, (2019, p. 16) we choose to avoid the 'hate speech' label in our work. This is because, despite its conceptual elasticity and the fact that the term itself does not feature explicitly in many countries' laws (Benesch, 2020, p. 13), the term 'hate speech' is too easily (even if incorrectly) associated with narrow legal categories of speech prohibited under national legal rules (Brown, 2017).[5]

4. Indeed, some platforms' 'hate speech' rules already cover groups that would not necessarily be considered protected by various countries' hate speech legislation (see Benesch, 2020).

We believe that in the platform governance context it is helpful, even if challenging, to maintain a distinction between this legally-defined category of content (hate speech which would be considered 'illegal' and therefore reasonably subject to stricter platform moderation in the form of takedowns) and another set of content that does not meet legal harm thresholds, but is nevertheless harmful (either individually or in aggregate) and may warrant platform intervention of a proportionate nature. Accordingly, and adopting a phrase popularised in the UK's Online Safety regulation debate, we conceptualise some forms of humour targeted at the historically marginalised as 'legal but harmful content'. This means we consider that this type of humour, as it is defined in this paper, may warrant platforms' intervention because it can be harmful. The intervention, however, need not be limited to takedowns and user bans, but can include other remedies, such as the elaboration of specific media literacy resources on humour and harm by platforms, similar to the educational materials some platforms published regarding misinformation during the COVID-19 pandemic (Ndiaye, 2021).

To develop our general argument on the importance of addressing humour for online safety, we use critical humour studies to explain why the particularities of humorous expression deserve special attention. To make our case for the relevance of addressing 'legal but harmful humour' targeted at historically marginalised groups, we draw on sociolegal and philosophical literature on harms (Bell, 2021; Gelber & McNamara, 2016; Gelber, 2019; Friedlaender, 2018; McGowan, 2009; McTernan, 2018), which recognise the *cumulative* nature of certain harms, as well as the pernicious yet frequently overlooked cases of harm that occur at a societal, rather than individual, level.

To help platforms in their elaboration of risk assessments and best practices regarding 'legal but harmful' content, this paper also suggests that Feminist Standpoint Theory's (Alcoff, 1991; Haraway, 1988; Harding, 1992; Collins, 1990) concepts of positionality and 'discursive context' can be useful additional tools for addressing 'legal but harmful' humour targeted at historically marginalised groups, and for differentiating this from legitimate and harmless humorous expression. The concept of positionality recognises that people's subject positions, from which values are interpreted and constructed, impact meaning and truth (Alcoff, 1991). It also

5. A variety of platforms' own terms of service can and do use the term 'hate speech' in broader ways than many countries' laws do (Brown, 2017), or in the case of platforms like Facebook, employ an expansive definition of 'hate speech' and break it down into further subcategories according to the severity of the speech and its harms. But there is real variation across platforms' policies when it comes to the definition as well as the scope of the term 'hate speech' (Benesch, 2020, p. 9). This is an additional factor behind our decision to avoid employing the term 'hate speech' as a broad descriptor for the type of content we deal with in this paper.

invites reflection on the 'discursive contexts' in which utterances take place, and how these contexts are aligned with, or resist, structures of oppression (Alcoff, 1991).

We divide the paper into three main sections. First, we do definitional work around humour – how it can harm in general, and in particular on digital media platforms when it is targeted at historically marginalised groups. Second, we discuss why the term 'online harm' has become central to the platform governance debate and situate humorous content and expression within this debate. Third, we explore positionality and 'discursive contexts' as conceptual mechanisms that can help social media platforms to explain and assess when lawful humour targeted at historically marginalised groups is likely to harm. The paper concludes with a discussion on the opportunities of taking humour seriously in regulatory efforts around online safety.

Humour, harm and digital platforms

The social function of laughter has received extensive scholarly attention, with critics proposing the superiority, incongruity and relief theories of humour – none of which are mutually exclusive – as answers to the question *why do people laugh?* The superiority theory views laughter as a means of dominating others; the relief theory understands it as a social 'safety valve' – people laugh to release tension; the theory of incongruity, meanwhile, recognises laughter as occurring when humorous expression does not adhere to logical expectations (Morreall, 1986). What makes different cultures laugh, authors argue, is a good indicator of their preoccupations and their relationships with power (e.g., Beard, 2014). The purpose of this paper, however, is not to assess why people might find different forms of lawful humorous expression on social media funny. Instead, our interest lies in unpacking the risks of individual and societal harm derived from lawful social media humour that punches down[6] on historically marginalised groups, and in stressing why we believe platforms should develop considered and proportional responses to this problem with appropriate nudging from media regulators. Relatedly, we are also interested in the risks of societal harm that arise when platforms erroneously take down critical and harmless humour. To this end, definitional work around what humour is and how it can harm historically marginalised groups is needed.

Humour operates according to a different set of rules than do other speech acts

6. Punching down is a term used mainly in stand-up comedy to describe the practice of ridiculing, parodying or mocking those with less privilege in society (Davis & Ilot, 2018).

(Morreall, 2009). First, it is highly ambiguous. In their humorous utterances, people often use diverse, overlapping and ambivalent rhetorical devices such as irony, sarcasm, hyperbole, satire and parody. The essential feature of satire, for example, is that it "aims to denounce folly and vice and to urge ethical and political reform through the subjection of ideas to humorous analysis" (Stott, 2004, p. 156). Satire can also include the use of irony, which is "the expression of one's meaning by using language that normally signifies the opposite, typically for humorous or emphatic effect" (Oxford English Dictionary, n.d.). Sarcasm, in turn, is similar to irony but less subtle and often used in a harsh tone or manner (Kreuz & Glucksberg, 1989). Humorous expression, then, often breaks typical conversational rules such as "avoid ambiguity" and "do not say what you believe to be false" (Morreall, 2009, p. 2). Second, the apparent 'non-seriousness' of humour explains why its negative consequences, or harms, have often been overlooked (Kuipers, 2011; Lockyer & Pickering, 2005). Critics note, however, that humour is a serious matter deeply entwined with power relations (Davies & Ilott, 2018, p. 6; Kuipers, 2011; Lockyer & Pickering, 2005).

Within deeply unequal societies in which certain groups experience historical and continued structural oppression, critical humour studies scholars assert that humour which punches down on these groups (e.g., racial and sexual minorities) can be socially corrosive because it stigmatises them (Kuipers, 2011; Lockyer & Pickering, 2005; Weaver, 2011). Research from psychology studies have also shown how lawful humour targeted at historically marginalised groups can cause individual psychological harm by silencing individuals from such groups and denying them equal social status (La Fave, 1977; Fry, 1977). Such humour has also been shown to increase communities' tolerance for discrimination and violence against the individuals belonging to these groups (Ford et al., 2008; Thomae & Viki, 2013). Importantly, overarching theoretical frameworks arising from the knowledge generated by historically marginalised groups (e.g., Critical Race Theory; Feminist Theory, Black feminist thought) have also recognised humour's potential to harm (Ahmed, 2017; Collins, 1986; Matsuda et al., 1993), arguing that in some cases, social and private sanctions (rather than the law) pose an "opportunity for success" in the regulation of 'legal but harmful' jokes (Matsuda et al., 1993, p. 43).

The ambiguity of humour and its potential to harm are exacerbated on social media platforms. Comedy sketches, parodies and satire on public social media are no longer only the work of media celebrities, artists, performers or journalists. Instead, everyday people have a platform to engage in different humour genres and forms. For example, while Blackface and Yellowface are theatrical traditions in

which largely white actors have portrayed Black and Asian characters in highly negative stereotypes, on social media, ordinary white users engage in these racist parodies as part of their everyday media practices (Matamoros-Fernández, Rodriguez & Wikström, 2022). The pervasiveness of these practices online has pushed some, but not all, social media platforms (e.g., Facebook) to ban Blackface and other racist stereotyping in their policies as 'harmful' content (Meta, 2022). Further, content on social media is also easily searchable, replicable, and scalable (boyd, 2010), affording humorous expression wider reach than originally intended, often in a decontextualised manner. Intent is also particularly difficult to assess on social media, since the folkloric ambivalence of humour is pushed into "hyperdrive" by the affordances of digital media, such as modularity, modifiability, archivability and accessibility (Phillips & Milner, 2017, p. 46).

Humour and its various rhetorical devices do feature, to varying degrees, in platforms' user-facing rules – often spread across different policies (see Appendix). But there are serious gaps in platform policies when it comes to explicitly addressing the limits of humorous expression. A comparative assessment[7] of the policies of platforms likely to meet or come close to meeting the EU Digital Services Act's definition of 'Very Large Online Platforms', or the UK Online Safety Bill's threshold for a 'Category 1' platform (Facebook, Twitter, YouTube and TikTok), reveals that Facebook's policies deal with humorous expression most comprehensively. Facebook's community guidelines make an implicit association between humour and vulnerability in various places (e.g., prohibiting the mocking of victims of sexual exploitation, those with a serious disability, and victims of hate crimes); Twitter, YouTube and TikTok, meanwhile, provide examples of 'hateful' and 'harmful' content that lean towards more overt forms of hate, leaving it unclear as to where humour that punches down on historically marginalised groups would sit. Moreover, Facebook, Twitter, YouTube and TikTok all make no clear distinctions between humour's rhetorical devices, such as satire, irony and parody. This matters because the four platforms give 'satire' and 'parody' special protection across various policies, but all fail to substantively define either (see Appendix for summary of plat-

7. We conducted our analysis of platforms' policies mentioning different types of humorous expression in mid-2021 and repeated the exercise in mid-2022. In just one year, Facebook, Twitter and YouTube had changed their policies regarding humorous content and conduct, especially in relation to misinformation. Since 2022, Facebook has a policy against mocking someone based on their protected characteristics who have COVID-19; Twitter's 'civic integrity' policy notes that in election time, even humorous/satirical content might be removed as part of efforts to limit misinformation; and YouTube's 'election misinformation' policy notes that content that violates this policy may be allowed if it includes additional context such as satirising misinformation. In 2022, Facebook also included a 'satire' exception across a number of policies, including hate speech, following recommendations from the Oversight Board's decision on the 2021-005-FB-UA case (Oversight Board, 2021).

form policies). Platforms also pay attention to 'intention' in their hate speech policies (e.g., incitement to violence), but do not explain how they treat ironic hate speech. Platform policies remain purposefully vague in order to provide a space in which interpretation can take place. But vague definitions around humour and its various rhetorical devices only creates more ambiguity around an already ambiguous concept. This, combined with a lack of recognition that lawful humour can play a role in cumulative harm to historically marginalised individuals and contribute to systemic societal harms, complicates platforms' efforts to set up effective governance processes that protect historically marginalised groups from abuse in a way that is balanced appropriately with platforms' freedom of expression obligations (Dias Oliva et al., 2020; Fielitz & Moore, 2018; Matamoros-Fernández, 2017; Paasonen et al., 2019).

The 'online harms' debate and humour

The term 'online harm' has become central to debates concerning both "regulation *of* and *by* platforms" (Gillespie, 2018). The EU and UK are just two of a growing number of jurisdictions (Linklaters, 2021) pioneering regulation with the stated aims of pushing platforms to protect free expression whilst also transparently and consistently addressing the spread of harmful content and behaviour. The challenge is that the notion of 'harm', much less 'online harm', is not self-evident, and important critiques have focused on how '[online] harm' is being conceptualised by platforms (DeCook et al., 2022) and by those pushing for measures to address it (Nash, 2019a; Turillazzi et al., 2022, p. 10). The question of definition and scope has proven to be particularly contentious[8] within jurisdictions like the UK, which wanted to regulate not only illegal content and conduct that causes harm (e.g., speech that harasses or incites violence), but also 'legal but harmful' content and activity. Critically, in the case of current definitional disagreements about '[online] harm', this is not some abstract, philosophical contestation – it is happening in the shadow of impending regulation where debates about the appropriate definition of 'online harm' have been tinged by concerns about the potential for regulatory overreach associated with more expansive or nebulous definitions (Bartolo &

8. The inclusion of the 'legal but harmful' category has been the subject of controversy. A number of UK-based civil rights groups, including Open Rights Group, Big Brother Watch and Legal to Say, Legal to Type have argued that the 'legal but harmful' category will significantly curtail free speech (particularly political dissent) and will be especially burdensome for smaller tech players that are unlikely to have the resources to comply (see Digital, Culture, Media and Sport Committee, 2022, p. 11). At the same time, a coalition of UK organisations argued that "[i]f done properly, the inclusion of legal but harmful content within the scope of this legislation could dramatically increase the ability for a wider range of people to exercise their free speech online by increasing the plurality of voices on platforms, especially from minority and persecuted communities" (Hope not Hate, 2021).

Matamoros-Fernández, forthcoming). In this paper, our focus is mainly on disputes and controversies related to the scope and definition of 'harm' within the '*legal* but harmful' category of emerging online safety regulation.[9]

There is a long history to the idea that the risk of individual 'harm' may warrant regulatory intervention (Nash, 2019b). In particular, liberal notions of 'harm' captured in the work of theorists like John Stuart Mill and Joel Feinberg are foundational in contemporary debates around legitimate regulation, including 'online harms' regulatory debates. Writing in nineteenth century England, philosopher John Stuart Mill (1859) posited "the harm principle" as the threshold that needs to be reached to justify intervention (including state intervention) into individual affairs. For Mill, individual liberty, including the liberty of expression, was paramount, and the state – as well as society more broadly – was only justified in encroaching on individual liberty "to prevent (the individual from causing) harm to others" (1978, p. 20). Because he viewed liberty of expression as fundamental to individual flourishing, Mill was only willing to concede the potentially harmful nature of speech acts in a highly narrow subset of cases: when speech constituted clear incitement to violence and defamation (see Riley, 2009).

A century after Mill, the American legal philosopher Joel Feinberg (1987) produced a widely cited four-volume treatise on harm. Whereas Mill's philosophical work on harm was concerned with delimiting a "personal zone of social non-interference" for individuals, Feinberg explicitly sought to theorise harm as a way of establishing the boundaries of the legitimate criminal punishment of individuals (Bell, 2021, p. 165). Feinberg defined harm as the wrongful "thwarting, setting back, or defeating of an interest" (1987, p. 33) and was particularly concerned with setbacks to individuals' "welfare interests". By "welfare interests", he meant the basic requirements necessary for individuals to build their version of a "good life". These included a range of interests from "the absence of absorbing pain and suffering"(physical harm) to "the capacity to engage normally in social intercourse and to enjoy and maintain friendships" (Feinberg, 1987, p. 37). Like Mill, Feinberg focused on incitement or threat as necessary conditions for defining speech acts as 'harmful', which has informed most speech legislation around the world (Sinpeng

9. The scope of the 'illegal harm' category has also been the subject of ongoing controversy, with criticism of earlier drafts of the Bill pointing out that it presented real dangers of overreach by seemingly adopting the term 'illegal' as defined by criminal, civil and administrative law (Digital, Culture, Media and Sport Committee, 2022, p. 8). The Bill presented to Parliament in March 2022 appears to partially address this concern by providing further specification, including a list of 'priority offences' (Schedule 7). Nevertheless, questions remain, including those regarding the Bill's apparent promotion of 'proactive technology' for platforms to detect illegal content. While important, these issues are beyond the scope of this paper.

et al., 2021). Yet various European countries' laws and regulations go beyond incitement to cover, for example, other harmful speech, such as "negative stereotyping or stigmatisation" (e.g., media regulation in the UK), "denying etc. acts of mass cruelty, violence, or genocide" (e.g., criminal statutes in Spain) and "dignitary crimes or torts" (e.g., criminal law in Germany and Switzerland) (Brown, 2015, chapter 2).

In the US context, an outlier in terms of speech regulation (Brown & Sinclair, 2020), Feinberg (1988) elaborated a complex "offence principle" to help legislators intervene on speech and conduct that, while not meeting the harm threshold, warranted (predominantly non-criminal) restriction, including some cases of racist jokes. Under his "offence principle", Feinberg (1988) proposed a set of criteria to evaluate the "seriousness" and "reasonableness" of offensive behaviour. Among other things, judging "seriousness" included evaluating the "extent", "duration" and "impact" of a particular offence. Judging a behaviour's "reasonableness" included assessing whether the behaviour in question had an important "social value" or constituted important speech worth protecting. The question of where to draw the line between speech that is 'harmful' and speech that is 'offensive', and between which of that offensive speech is permissible and which non-permissible, is, however, hotly contested. This is particularly well-demonstrated in the Western European context, where greater legal constraints are placed on harmful speech as compared to the US (Kahn, 2013), where "offensive expression" has also faced restriction (O'Reilly, 2016).[10] In Western Europe, the European Court of Human Rights (ECtHR) has invoked the right of expressions to 'offend, shock or disturb' in several humour-related cases, with the aim of striking a balance between restricting humour when it is 'gratuitously offensive[11]' and protecting humour when it contributes to 'public debate' or has a 'public interest' value (Godioli, 2020; Kuhn, 2019). In very few cases, the Court has prioritised the protection of individuals' human rights (e.g., the right to not be discriminated against) over humour's 'public interest' value (e.g., Féret v. Belgium, 2009). As Godioli and Little (2022) note, de-

10. Legal scholar Robert A. Khan (2013) argues that "[t]he Euro-American debate over hate speech laws has been ongoing and more varied than one might expect" (p. 552) and he notes that "one can see a change from the 1930s-a time when speech restrictions appeared to be the modern, democratic wave of the future, and the current situation, in which, at least from an American perspective, the converse appears to be true" (p. 585).

11. The European Court of Human Rights (ECtHR) deems expression to be "gratuitously offensive" (and therefore not necessarily worthy of protection) if that speech has "no basis in fact or [...] is needlessly insulting" (O'Reilly, 2016, p. 241). The controversial "gratuitously offensive" test was introduced in Otto-Preminger-Institut v. Austria (1994), in which the ECtHR ruled that the Austrian government's censorship of a satirical film targeting Catholics did not violate the right to freedom of expression under Article 10 of the European Convention on Human Rights.

spite the central position of dignity in Western European approaches to free speech regulation, the ECtHR has been inconsistent in using harm as an objective test to determine when humour is unlawful, turning instead to the 'gratuitous offence' test for restricting humorous speech.

Further, conflating harm with offence is all too frequently a strategy used to undermine the case for forcefully addressing speech that harms, especially speech targeted at historically marginalised groups (Waldron, 2012, p. 111). Beyond (if inseparable from) the strictly legal domain, there is an important sociopolitical dimension to ongoing global debates over the permissibility of 'offensive' speech, including in Europe. Especially since the 1990s, in a context of rising global anxieties about multiculturalism, LGBTIQ+ rights and Islam, the language of 'liberal (free speech) values' and the 'right to offend' has been strategically deployed by some to further their populist agendas with real costs for social integration (Larsen, 2013; Maussen & Grillo, 2014; Rostbøll, 2008).

In the US context, legal scholar Jeremy Waldron (2012) offers an important critique of using 'offence' to assess the permissibility of speech, insisting that when addressing the harms of speech targeted at historically marginalised individuals and groups, regulators need to consider its drip-drip effects, not just its immediate ones. He goes on to argue that some forms of abuse targeted at historically marginalised groups do not merely 'offend' their targets by evoking "subjective [...] hurt, shock and anger" (p. 106), but they objectively harm them by attacking their dignity, understood as their social standing, which is more aligned with how Western European legislation conceptualises speech *harms*. Similarly, also in the US context, philosopher Melina Constantine Bell (2021) rejects the 'offence principle' for regulating 'low key' abuse targeted at historically marginalised groups. She argues that "modern scientific knowledge" about the connections between bigoted speech and "forms of tangible psychological harm, mechanisms for transmitting cultural norms, implicit bias, structural discrimination, etc.", would indicate that this speech, including humorous speech, is more appropriately assessed through a 'harm' lens (p. 163). Waldron's and Bell's work retains its relevance in the European context. Although it has long been accepted (not unreasonably) that there is a fundamental difference, even incompatibility, between the US and Western European approaches to regulating speech. The "European consensus" on banning harmful speech, such as hate speech, has come under strain in the twenty-first century and "settled assumptions" about restricting "provocative expression" are no longer quite so settled (Heinze, 2013, pp. 591-592).

One of the philosophical principles underpinning the need for expansive conceptu-

alisations of speech-related harm is what political and legal theorist Alexander Brown calls "the principle of non-subordination", which holds that speech restrictions are justified "if they serve to protect individuals from acts of expression that also constitute acts of subordination" (Brown, 2015, p. 75).[12] This principle builds on Critical Race scholarship by authors such as Richard Delgado and Jean Stefancic, and feminist philosophical thought from scholars such as Rae Langton, Mary Kate McGowan, Ishani Maitra and others). Crucially, drawing on their "situated knowledges" (Haraway, 1988), these scholars claim that speech 'does things'[13] and therefore it can both cause and constitute harm. This provides a potent reminder of the limits of holding up dominant liberal conceptualisations of harm, including those of Mill and Feinberg, as socially neutral. Subsequent work has unpacked the various ways in which, by drawing definitional boundaries around the 'harm' concept to satisfy their own particular worldview, liberal theorists like Mill and Feinberg were able "to conceal real injuries, and to marginalise some conceptions of the good life" (Smith, 2006, p. 3).

Another especially valuable insight from critical race and feminist scholars has been detailed theorisation of the ways in which the harms of speech not only connect with, but are inseparable *from*, broader contexts of structural, social and political inequality and oppression (e.g., see McGowan, 2009; McTernan, 2018). This contextualisation is particularly crucial when dealing with various forms of 'low level' or subtly discriminatory speech, including certain types of humorous expression targeted at historically marginalised individuals and groups. This speech would appear to be (and may in fact be) less severe than the kinds of speech acts that meet criminal thresholds, yet its potential to harm becomes evident once it is assessed against a "background condition" of structural injustice (Friedlaender, 2018). Indeed, one powerful critique of liberal conceptions of harm proposed by theorists like Mill and Feinberg has been their tendency to centre individual actions and agency over structural factors when deciding whether a harm threshold has been reached (e.g., Pemberton, 2015). And yet, for individuals belonging to groups experiencing historical and continued structural oppression, "routine" and "subtle" forms of abuse have "cumulative effects" (Gelber & McNamara, 2016, pp.

12. To subordinate someone is "to put them in a position of inferiority or loss of power, or to demean or denigrate them" (Langton, 1993, p. 35).

13. The idea that speech 'does things' comes from speech act theory (J.L Austin, 1962), which Rae Langton, Mary Kate McGowan and Ishani Maitra draw on to conceptualise some forms of speech as both causing and constituting subordination when uttered with authority. Influential previous work on the harms of speech–feminist work by Catharine A. MacKinnon and Andrea Dworkin, and critical race theory work by Mari J Matsuda, Charles Lawrence, Richards Delgado and Kimberle Crenshaw – does not draw on speech act theory. However, all these authors coincide in claiming that speech can do harmful things (for a detailed review of this literature see de Silva, 2020).

500-501; see also Freeman & Schroer, 2020; McGowan, 2009). In the case of humour, the idea of cumulative harm is specifically pertinent due to the relevance of "joke cycles" – patterns in the use of humour that repeat over time in the public sphere (Ellis, 1991). These cycles should not be dismissed: as philosopher Emma McClure has written, "jokes and threats [were] mixed together" in the lead-up to the Rwandan Genocide, yet the significance of the 'jokes' only became clear in hindsight, once the violence had started (McClure, 2020, p. 132). But our central point here is that even when demeaning jokes are *not* followed by such horrifically brutal events, they matter to the extent that they contribute to cumulative harm for individuals; further, such jokes both indicate and perpetuate forms of structural inequality that simply cannot be measured 'merely' by acts of overt violence.

It is important to specify that standards of evidence for 'harm' differ when discussing individual-level versus societal harm. As legal scholar Nathalie Smuha (2021) has argued, societal harms often occur over the longer term, making it difficult (and not necessarily possible) to prove a direct causal link between a specific act and a harm. As she notes, one area where the individual-level harm framing has been successfully challenged is environmental law, where it has been recognised that harms such as pollution accumulate over time, have distributed effects across society and require a different framing to that of individual-level harm. In this sense, overly individualised frameworks for conceptualising 'legal but harmful content' (as currently envisaged in the UK's Draft Online Safety Bill, for example) may prove limiting. The case for attending to societal-level harm on social media has been most strongly made (although not at all uncontroversially) in cases of mis/disinformation, matters of public health and national security,[14] as well as in relation to new threats to privacy introduced as a result of practices such as group-level algorithmic targeting, which have effects far beyond any specific individual (e.g., see Mittelstadt, 2017; Smuha, 2021). And yet there is also a case to be made for adopting a societal harm lens to assess and mitigate low-level online abuse targeted at historically marginalised individuals and groups.

Like the 'legal but harmful' category, the 'societal harm' concept is open to the charge of being too nebulous, too easily seized on by those in power to justify all forms of regulatory overreach (see Joint Committee on the Draft Online Safety Bill, 2021, p. 36). Yet, sensitivity to contextual structural and societal dynamics is criti-

14. For example, this is one of the ways in which VLOPs are thought to contribute to potential 'societal risk' in the EU's Digital Services Act. In the UK context, academics such as Democracy scholar Alan Renwick (2021) have argued that harms to democracy should be incorporated into the Online Safety Bill, and the overly individualised nature of the Bill's conceptualisation of 'harm' has also been lamented by others (e.g., George, 2022; Edwards, 2021).

cal when it comes to assessing and proportionately addressing online harms. A sophisticated approach to 'online harms' invariably necessitates dealing with questions of how online content and behaviour dynamically interact with other factors to produce effects in a rapidly changing, interlinked and convoluted world (Hargrave & Livingstone, 2006). In the case of online humour that punches down on historically marginalised groups, platforms should recognise its harms when this humour is assessed within the context of pre-existing structural oppression (e.g., police violence against racialised people, hate crimes against transgender people) (Bell, 2021). This then raises the question of what a nuanced, proportionate platform response might look like to minimise humour's (risk of) harm.

Solutions to online harms

An important implication of defining 'harms' more expansively beyond the confines of existing legal frameworks and taking a systemic approach to identifying harms beyond individual pieces of content and user violations is that this could (and should) also give way to the implementation of diverse remedies beyond content takedown and user bans, which are themselves inspired by the existing criminal justice system and have historically dominated platform content moderation approaches (see Goldman, 2021; Schoenebeck et al., 2021). In early 2022, a House of Commons Committee Report (Digital, Culture, Media and Sport Committee, 2022) recommended that the UK's Bill be amended to "include non-exhaustive, illustrative lists of preventative and remedial measures beyond takedowns for both illegal and 'legal but harmful' content, proportionate to the risk and severity of harm, to reflect a structured approach to content" (p. 26). The Report provides examples like "tagging or labelling, covering, redacting, fact-checking, deprioritising, nudging, promoting counter speech, restricting or disabling specific engagement and/or promotional functionalities (such as likes and intra- and cross-platform sharing) and so on"(p. 26). In earlier versions of the UK's Bill 'Category 1' platforms were required to specify in their risk assessments and terms of service how they would deal with legal but harmful content (Section 13(3)). Platforms would also be required to provide so-called "user empowerment tools'' to enable users to limit their exposure to content they choose not to engage with (Section 14).

One would hope that platforms' risk assessments and terms of service will make clear how platforms' proposed remedies relate to their judgement not only about the severity but also about the *nature* of a particular harm (Benesch, 2020, Proposal 5). For example, "user empowerment tools", like the ability for an individual user to change the settings of their personalised recommendations or 'unfollow' an ac-

count, may help to shield individuals from certain harms. But these tools are woefully inadequate when it comes to dealing with many harms of a more *societal* nature (Milano, Taddeo & Floridi, 2020) and can also place an excessive burden on those most affected by online harms, who end up doing continuous "safety work" to protect themselves (Gillett, 2020). Determining which actions are appropriate to deal with particular (risks of) harm adequately requires platforms to tailor their assessments around proportionality, associated trade-offs and overall effectiveness. There also remains crucial (theoretical and empirical) work to do in the case of 'beyond removal' remedies like content downranking, interstitial warnings, reduced engagement functionalities, etc.[15] We intend to contribute to this important work through future publications.

Nevertheless, before determining the most adequate remedies for content and conduct that can harm, social media platforms still need to determine first *when and how* content (individually or in aggregate) is likely to harm. We address this question in the next section by focusing on the difficult task of understanding and assessing the limits of humour, and we suggest that concepts and principles from critical humour studies and Feminist Standpoint Theory can help platforms in this endeavour.

Using positionality as an additional tool to assess when humour is likely to harm

Evaluating when humour can harm is not an easy task, as evidenced by research that has studied how courts struggle with assessing cases involving humour (Godioli, 2020; Godioli & Little, 2022; Little, 2008). Humour related jurisprudence in both the US and Western Europe shows that courts have a limited understanding of the different rhetorical devices used in humorous utterances, which has undermined their ability to consistently and proportionately assess this form of expression (Godioli & Little, 2022). Godioli and Little explain how courts often use the terms satire, parody and humour "interchangeably", which leads them to interpretive problems when assessing humour (p. 308), a terminological inconsistency we also observed in our analysis of social media platforms' policies. They argue, then, that insights from critical humour studies "can set the basis for a more fine-grained and systematic approach to humour across different judicial systems" (p. 305). Digital platforms, as important actors making decisions around humour, could also

15. But see, for example: important theoretical contributions to the 'beyond removal' debate by Tarleton Gillespie (2018), Eric Goldman (2021), Blake Hallinan (2021), Paddy Leerssen (2021) and Luke Munn (2020) amongst others.

benefit from insights drawn from critical humour studies. This is because this scholarship not only has unpacked how humour operates according to a different set of rules than do other forms of expression (Godioli, 2020; Godioli & Little, 2022; Little, 2008), but also because it sheds light on the significance of structural factors, such as the speaker's power when assessing the limits of humour.

In fact, scholars across diverse disciplines have stressed the importance of considering structural factors in determining the likelihood of speech to harm, including humorous expression. Humour scholars have noted that the responsibilities of 'authors' of humorous texts cannot be subordinated to "the ethics of reading" (Davies & Ilott, 2018, p. 15; Kuipers, 2011). That is, as Davis & Ilot (2018, p. 15) explain, the reception of humorous expression in public spaces "is all-too-often collapsed into polarising debates around free speech, intentionality, and offence, in which the artist's [in the case of social media, everyday users in general] right to free speech is often held sacred above even the sacred as conventionally understood" (Davis & Ilot, 2018, p. 15). However, they argue, less attention has been given to the speaker's power when assessing humour's potential to harm (Davis & Ilot, 2018). The notion of the speaker's power in this scholarship is closely linked to Bakhtin's (1982) concept of the "carnivalesque", a satirical tradition in which members of lower classes use humour to subvert oppressive power structures.[16] This conceptualisation is a helpful starting point for connecting humour to questions of power, and for identifying cases in which humour may be harmful in particular instances (e.g, when differentiating between genuine satire, which is socially *corrective* because it punches up at figures with 'authority'[17] (e.g., political elite), and 'pseudo-satire', which is socially *corrosive* because it performatively exploits sarcasm and irony to punch down on vulnerable groups) (Mondal, 2018). Yet using this approach to understanding the speaker's power in humorous exchanges can fall short. By adopting a focus on class/social status (e.g., proletariat vs. bourgeoisie), other dimensions of power (e.g., along dimensions of race, gender and their intersections) are often not exhaustively treated. Scholars like Davis & Illot (2018) do lay important groundwork in urging scrutiny of *"our various subject positions [...]* from our *multifarious sites* of *privilege and/or subordination"* when consid-

16. Notions of power in Bakhtin's carnivalesque tend to revolve around general relations of domination and subordination within feudal culture in the Middle Ages – processes of "liberation from oppressive norms […] 'from the prevailing truth and from the established order'" (Stevens, 2007, p. 1). Contemporary humour studies scholars like Simon Critchley (2002) and Davis & Ilot (2018) repurpose Bahtkin's carnivalesque in their highlighting of humour/comedy's potential to "challenge the status quo and to give agency and expression to the disempowered", as well as its capacity to reinforce hegemonic power structures (Davis & Ilot, 2018, p. 9).

17. See our later discussion on the concept of authority.

ering the impacts of humourous exchanges (p. 16, our emphasis), yet leave plenty of room for more systematic and nuanced theoretical approaches based on frameworks designed specifically for treating intersectionality and positionality.

Feminist philosophers of language have also long argued that the speaker's "authority" is crucial in determining whether or not speech-acts can be harmful (Langton, 1993; Maitra, 2012; McGowan, 2009, p. 389). Generally, people have authority if their formal designations allow them to "assign tasks to others" or to influence norm-setting (e.g., a teacher or a legislator – Maitra, 2012, p. 14; see also Langton, 1993). Authority, though, can also be "granted" (Maitra, 2012, p. 107): a speaker without a recognised position of authority can have "derived authority" if those who have formal authority delegate that authority to that speaker or fail to intervene on that speaker's speech/actions (Maitra, 2012, pp. 104-105). Authority can also emerge from the social context; namely speech is likely to be harmful if what is being said conforms with, for example, systems of gender and racial oppression (McGowan, 2009).

Concepts from Feminist Standpoint Theory such as "positionality" and "discursive contexts" (Alcoff, 1988; 1991) add nuance to previous theories of the role played by structural factors in determining when speech, including humorous expression, is likely to harm. Specifically, the concept of positionality is particularly relevant for how it draws attention to the ways in which people's diverse and intersecting subject positions[18] influence what speech does to others and affects "the meaning and truth" of what is being said (Alcoff, 1991, p. 6). Positionality (e.g., Alcoff, 1988; 1991) invites reflection on the effects that everyday interactions have depending on the multiple subject positions people occupy (not only according to their socioeconomic status or formal authority, but also how they are situated within what Patricia Hill Collins (1990) calls "the matrix of domination"). This means that, for example, a middle class white queer woman from a Western country occupies a position of privilege in terms of her race and class, but her position within a network of gender relations lacks power relative to white cis gender men in a patriarchal society. At the same time, Black women have historically experienced simultaneous and cumulative oppressions due to the interlocking nature of their subject positions in terms of race, gender, class and sexuality (Collins, 1990; Crenshaw, 1991). Similar to the arguments put forward by feminist philosophers of language, the external contextual factors within which people are situated influence people's

18. We use 'positions' in plural because our positionality, like our subjectivity, is fluid and needs constant self-evaluation and reflection. Alcoff (1981, p. 7) also uses 'social location' and 'social identity' to refer to speakers' subject positions.

relative positions to others, "just as the position of a pawn on a chessboard is considered safe or dangerous, powerful or weak, according to its relation to the other chess pieces" (Alcoff, 1988, p. 443).

In her essay about the problems involved in speaking for others,[19] Latin American feminist philosopher Linda Alcoff explains that to understand discursive practices of speaking or writing (what she calls 'rituals of speaking'), one should focus on what is being said/written in combination with paying attention to the speaker's positionality and the broader context (1991, p. 12). Alcoff writes: "Who is speaking to whom turns out to be as important for meaning and truth as what is said; in fact, what is said turns out to change according to who is speaking and who is listening" (1991, p. 12). For her, a speaker's location not only is "epistemically salient", but "certain privileged locations are discursively dangerous" (p. 7). In particular, Alcoff maintains that specific practices of privileged people, such as speaking for or on behalf of less privileged people[20] (e.g., in our case, for example, tech and political elites defining 'online harms' without consulting those most affected by online abuse), have the potential of "increasing or reinforcing the oppression of the group spoken for" (p. 7). We consider that other practices of privileged people, such as making fun of those with less privilege in society, can also reinforce systems of oppression. Like positionality, the "discursive context" in which utterances are made is also crucial for unpacking 'rituals of speaking'. Alcoff defines the discursive context as "the connections and relations of involvement between the utterance/text and other utterances and texts as well as the material practices in the relevant environment, which should not be confused with an environment spatially adjacent to the particular discursive event" (p. 12). She explains that certain contexts and locations "are allied with structures of oppression", while others "are allied with resistance to oppression" (p. 15). Therefore, she argues, not all contexts and locations are "politically equal, and, given that politics is connected to truth, all are not epistemically equal" (p. 15).

In the remainder of this section we describe two examples of what we consider 'legal but harmful' humour online targeted at historically marginalised individuals and groups, and show how being attentive to 'positionality' and 'discursive contexts' could help platforms recognise that the humour presented in these cases is likely to harm (rather than merely offend) and hence requires their intervention.

19. Alcoff describes the practice of speaking for and about others as "the act of representing the other's needs, goals, situation, and in fact, *who they are*" (p. 9).

20. Examples of problematic speaking for others, according to Alcoff (1991), are "the U.S. government's practice of speaking for [...] Third World nations" (p. 8) or privileged academics "assuming" the identity of less privileged people by writing semi-fictional characters in first person (p. 5).

This intervention, however, should be proportionate to the risk of harm associated with the content, meaning platforms should be willing to experiment with various remedies beyond the blunt tool of content removal.

Harmful parodies

Parody is a specific type of humour that is "forced to reference that which it mocks" (Davies & Ilott, 2018, p. 12). In 2020, African American TV writer, producer and actress Franchesca Ramsey called out YouTube on Twitter for having allowed its biggest creators to get rich on the platform at the expense of using minorities as the butt of these Youtubers' jokes (Ramsey, 2020). The tweet came as a response to YouTube creator Shane Dawson's apology for having performed Blackface in some of his past sketch comedy videos. Dawson's apology followed another video apology by YouTube star Jenna Marbles, who had parodied Trinidadian-born rapper Nicki Minaj in one of her early videos. Both Dawson and Marbles apologised in a moment of collective reckoning in the US (and globally) on race and racism following the 2020 Black Lives Matter (BLM) protests in the aftermath of the police killing of George Floyd. In the midst of this global (see Mendes, 2021) and national-cultural audit, these racist parodies were considered by some to be more socially unacceptable than they were in the past, while for others Blackface has always been deemed harmful (Jackson, 2019; Lott, 1992). The harm of Marbles' and Dawson's blackface performances and other racist parodies rooted in historic oppression is verifiable with access to relevant information: the parodies were made in public and targeted members of historically marginalised groups, and the speakers acted from a position of race privilege that can be considered "discursively dangerous" (Alcoff, 1991, p. 7) in the context of systemic racism in the US (Feagin, 2006) and other Western countries (Wolfe, 2016). Following Alcoff's call to consider positionality for evaluating the meaning of interactions, the details of who is parodying whom becomes integral to objectively determining the risk of harm derived from this type of humour.

While Facebook prohibited Blackface in 2020, YouTube and largely all other major social platforms (e.g., Twitter, TikTok) do not specifically prohibit this and other harmful stereotypes of historically marginalised groups in their policies. In Marbles' and Dawson's case, it was the YouTubers themselves who chose to engage with the 'apology genre' on Youtube (see Wolsey (2020) for a detailed examination of apology videos for racist acts) to publicly acknowledge they were punching down on historically marginalised groups. Without a clear recognition in its policies that negative stereotypes of historically marginalised individuals and groups can be harmful, and that structural factors are crucial in determining content's

likelihood to harm, YouTube remains unwilling to consider appropriate remedies to address legal but harmful content, such as racist abuse uttered in the form of parodies. In this case, the controversy emerged and was resolved without the need for specific policies, largely due to the context of increased awareness of systemic violence towards Black Americans in the US, pushing Dawson and Marbles to issue their apology. However, oppression of Black Americans in the US is systemic and was ever present when Dawson and Marbles performed Blackface years before 2020's BLM protests. Had YouTube developed clear policies regarding humour and its connection to harm, the platform would have been able to intervene in other controversies involving humour targeted at historically marginalised groups that were not organically resolved (see Roberts (2016) for examples of failed content moderation of humour that can harm on YouTube).

There are other cases of racist humorous expression on social media in which the balance between harm and freedom of expression is less clear-cut than in the racist parodies presented above. For example, in 2018, a caricature[21] by Australian cartoonist Mark Knight of American professional tennis player Serena Williams circulated on Facebook and Twitter. The cartoon, which attracted global attention, criticised Williams for snapping at the umpire when she lost against Naomi Osaka in the 2018 US Open women's final. Knight's caricature resembled racist illustrations during the US Jim Crow era and Sambo cartoons, which have a history attached to violence (Lott, 1992), and tapped into the misogynoir trope of the "angry black woman" stereotype (Ashley, 2014). Indeed, Mark Knight's cartoon could be considered an example of humour that punches up and down simultaneously. It punches up because it denounces poor behaviour from a professional athlete. However, by tapping into the "The Sapphire" stereotype, Knight's cartoon of Williams also becomes a racist and sexist portrayal of a Black woman by a white male cartoonist (Gatwiri, 2018). As we have argued in the previous sections, scholars have long argued that even though humour lends itself to multiple interpretations, when one of these interpretations reveals long-running racist and sexist tropes, for example, this humour should be carefully assessed for its potential to harm (Weaver, 2011). In addition, the fact that Mark Knight drew this cartoon from a position of race and gender privilege further complicates the justification of this type of humour as merely performing a social critique of a famous athlete, and as such warranting protection. We are aware that journalists are granted special protection in speech regulation in most liberal democracies. However, we use this example to show how when this and similar cartoons are circulated on social media

21. The cartoon was originally posted in *The Herald Sun*.

platforms, including in the European context (e.g., Oversight Board, 2022), the responsibility to judge this form of legal but harmful humour falls to platforms. There is no evidence of how Facebook or Twitter moderated this cartoon, but we consider Mark Knight's caricature of Williams to be another example of 'legal but harmful' humour that big social media platforms could address in their online safety efforts.

TikTok anti-Asian memes during the COVID-19 pandemic

Anti-Asian TikTok memes created during the COVID-19 pandemic are another instance of lawful humour that is likely to harm. Matamoros-Fernández et al. (2022) observed the salience of 'Yellow Peril' memes on TikTok during the early stages of the pandemic. 'Yellow Peril' is a long-running Western racist trope that has historically "played a crucial role in the cultural production of Asians as a racial contagion" (Mallapragada, 2021, p. 279), especially as it represents Asians as savages, merciless, immoral, subhuman, and a threat to [...] whites in general" (Ono & Pham, 2009, p. 38; see also Odijie, 2018, for a description of 'Yellow Peril' in a European context). During COVID, users on TikTok (mostly white people) engaged in the 'Yellow Peril' trope via the creation of highly popular[22] videos that targeted people of Asian descent as being the cause of coronavirus (Matamoros-Fernández et al., 2022). In these videos, people reacted in disgust and fear when they received parcels from China, and commonly engaged in humorous videos that attributed the origin of the pandemic to Chinese people's eating bats (Matamoros-Fernández et al., 2022).

These TikTok viral trends happened at a time of increased anti-Asian sentiment worldwide, which led to real-world violence (e.g., Gover et al., 2020; Gray & Hansen, 2021; Kamp et al., 2021). Whilst large platforms were vocal about the extra efforts being put into addressing the spread of health misinformation during COVID (Newton, 2020), there was no coordinated effort by platforms to provide extra information on whether they were stepping up their interventions to tackle the spread of racist content despite health authorities insisting that stigmatisation of people or groups (based on their ethnicity or the fear that they may have the virus) could have material consequences (IFRC, Unicef & WHO, 2020). While Twitter updated its 'hateful conduct' policy to prohibit hateful conduct on the basis of 'disease' in 2020 – though they made clear that this development was not triggered by COVID-19 in particular (Culliford, 2020) – TikTok did not change its Terms of Service, nor did it release any public statement in response to the increased stig-

22. The videos discussed by Matamoros-Fernández et al. (2022) received thousands of likes on TikTok.

matisation of people of Asian-descent during the pandemic.

The likely risk of harm in anti-Asian memes on TikTok is verifiable when taking into account relevant contextual information, such as the positionality of those engaging in these practices (largely white users acting from a position of race privilege) as well as the larger "discursive contexts" in which these practices took place (e.g., a broader context of global Sinophobia during COVID and some Western countries' long history of racism towards people of Asian descent – see, for example, Mallapragada, 2021; Odijie, 2018). In addition, scholarship around the racialisation of illness points to the immediate and longer-term social damage that attaching an illness to a particular racial group can inflict in multicultural and multiracial environments (e.g., Keil & Ali, 2006). Despite this contextual information being available to TikTok, the platform did nothing to address these anti-Asian memes. TikTok has, however, moderated other instances of controversial humour likely to harm in the past, albeit in an ad hoc manner (BBC News, 2020).

Some tech companies, such as Facebook, have experimented in the past with taking power relations into account when moderating content that is likely to harm. For example, as part of the WOW project ('worst of the worst'), Facebook weighted anti-Black, anti-Muslim, anti-Semitic and anti-LGBTQ hate speech as higher priority than hate speech directed at men or white people (Dwoskin, Tiku & Kelly, 2020). These initiatives to remediate online harms, though, are often impossible to audit from the outside, which means that platforms' self-governance efforts lack the appropriate systems for meaningful transparency and due process. Further, it is not clear that current approaches to content moderation, including automating the removal of harmful content/conduct, should be the only or the primary solutions to mitigating online harms (Douek, 2022; Duguay et al., 2020; Gillespie, 2020).

Our point being: we believe that some new online safety proposals that seek to create clearer guidelines around how platforms deal with online harms (e.g., the DSA) could be an opportunity to reconsider how and when humour targeted at historically marginalised groups can harm. We acknowledge that moving beyond legal frameworks to conceptualise 'online harms' invites legitimate fears of regulatory overreach. Yet we believe that the non-prescriptive procedural accountability/risk assessment approach being adopted by the EU could (at least in theory) help to partially alleviate such concerns. Additionally, regulators and platforms should pursue ongoing participatory initiatives to listen to the needs and interests of those most impacted by online harms (Schoenebeck & Blackwell, 2021, p. 14). This would limit those with the most political power (e.g., governments; private companies) from dominating the definitional boundaries of 'online harms', and in this way

help them to avoid speaking for others (see Alcoff, 1991) on the matter. Consultation processes could result in platforms (1) implementing more elaborate definitions of humour's various rhetorical devices in order to better assess this form of expression's risk of harm, and (2) clarify how structural factors influence the differential impacts of humour. This latter claim could prompt platforms to broaden the categories of content they currently address in their terms of service to include some forms of 'legal but harmful' humour targeted at historically marginalised groups that are likely to harm.

Although it remains to be seen how regulatory frameworks like the EU's will develop and be applied in practice, as 'system and process'-based regimes, they hold promise insofar as they may also prompt platforms to test (and make auditable) whether and how their own design and processes incentivise or enable online harms. For example, platforms could audit whether their automated approaches to content moderation (including their moderation of humour) over-identify, or fail to protect, those most affected by online harms, as current research suggests (Buolamwini & Gebru, 2018; Dias Oliva et al., 2021; Paasonen et al., 2019). These new regulations may also push platforms to seriously assess the risks associated with a steady build up of unsavoury (though not individually harmful) content/conduct, including some forms of humour, and to tap into a broader suite of proportionate remedies to address the wide variety and complex manifestations of online harms, beyond content removal and user bans. To be sure, while improved policies to remediate online harms are desirable for their "expressive value" – they signal "the standards of the community" and may help in "limiting the potential for future harm"– they also generate "new challenges in ensuring accountability in platform policy enforcement" (Marinett, 2021, pp. 10-24). For example, considering structural factors such as the speaker's power to assess the limits of humour does not mean that there will be a perfect, one-size-fits-all method to assess humour's risk of harm on social media. In fact, it is most likely that platforms' moderators[23] and their specialised content moderation teams could only operationalise positionality and 'discursive contexts' on a case-by-case basis. Scholars and advocacy groups have long voiced concerns about platform accountability when it comes to 'blunt' instruments for content moderation (e.g., takedowns/user bans) (e.g., Gillespie, 2018; Suzor, 2019). These concerns are likely to be exacerbated in the case of 'soft moderation' techniques (e.g., downranking, interstitial warnings) which are difficult to track/audit. Whilst remedies like downranking may help to limit the over-

23. There are studies that maintain that annotators' positionality influences the moderation of content (e.g., Larimore et al., 2021), which suggests that training which is sensitive to structural factors might be needed for those involved in platforms' content moderation processes.

removal[24] of content, it must be acknowledged that part of the challenge in the next few years will be to develop accountability mechanisms for these 'soft moderation' techniques (see e.g., Leerssen, 2020; Rieder & Hofmann, 2020).

Conclusion: Humour must be taken seriously to ensure online safety and wellbeing

In this paper, we have argued that policymakers and platforms should take humour seriously in their online safety efforts in order to protect historically marginalised groups. We have maintained that lawful humour targeted at historically marginalised groups can cause individual harm via its cumulative effects and also contributes to broader societal harms. Given the inherent challenges involved in drawing causal links between online content and specific harms (Hargrave & Livingstone, 2006), the importance of conceptualising and assessing harms in light of society's broader structural conditions cannot be overstated. Lawful humour that punches down on historically marginalised groups can harm (rather than '*just*' being subjectively offensive), which social media platforms could (when possible) objectively verify by accessing relevant situational information, such as factoring in broader systems of structural oppression and taking into account both the speaker's power and the discursive context in which humour occurs.

Our argument has been that expansive conceptualisations of 'harm' within online safety regulation present an opportunity to hold platforms accountable for how they deal with diverse online harms – including how they deal with humour that can harm historically marginalised individuals and groups via its cumulative effects and for its larger social implications. In this regard, by better understanding humour, platforms could more effectively address lawful content that is *already* prohibited in their Terms of Service, but hard to recognise due to its playful nature. Yet this paper is also an invitation for platforms to enlarge the scope of what they consider harmful to include some forms of lawful humour targeted at historically marginalised groups that are likely to cause social and, in aggregate, individual harm.

There has been varied and vocal resistance to the idea of addressing 'legal but harmful' content as part of online harms regulation (e.g., Index on Censorship, 2021; Schmon, 2021). Some of this resistance, though, can be resolved once it is acknowledged that platforms have many different tools available to them when

24. "Over-removal" is a term used in platform governance literature to describe how social media platforms take an "if in doubt, take it down" approach (Keller, 2015) to requests for takedowns (from governments, for example).

attempting to mitigate harm – tools which extend well beyond the "leave up/remove binary" that regulators have been most focused on (Goldman, 2021). This might include educating users by tagging or labelling content, or reducing certain content's visibility through algorithmic de-prioritisation or the restriction or disabling of engagement functionalities. Deepening our understanding of how those who bear the brunt of harmful content would like platforms to remedy this harm and developing a harm mitigation toolkit that is reflective of these voices is key (Schoenebeck et al., 2021). Yet developing effective remedies to problems must begin with an accurate diagnosis of those problems.

There is nothing preordained, objective or fixed about how a society chooses to define 'harm'; in fact, it is arguable that shifting definitions of 'harm' act as a barometer of our "dynamic cultural values" (Gardner, 1989, p. 8). They also act as a barometer of power. It is significant, for example, that policymakers have recently shown a willingness to evolve their understanding of online harm to include 'social harms' when it relates to activity that threatens to disrupt political power, and specifically platform-enabled disinformation that threatens to undermine state-based democratic systems. A focus on humour targeted at historically marginalised individuals and groups provides a glimpse into what it might look like if platforms and policymakers were to adopt the same willingness to evolve conceptualisations of harm in relation to activities that affect those with the *least* political power.

ACKNOWLEDGEMENTS

The authors would like to thank Alice Witt and Nicolas Suzor, and the three reviewers who read this paper for their very helpful comments and suggestions. Special thanks also go to the journal's editors Balázs Bodó and Frédéric Dubois for their valuable feedback. Matamoros-Fernández would also like to thank Amy Johnson for introducing her to the idea of positionality as a potential tool to assess humour and power.

References

Ahmed, S. (2017). *Living a feminist life*. Duke University Press. https://doi.org/10.1215/9780822373377

Alcoff, L. (1988). Cultural feminism versus post-structuralism: The identity crisis in feminist theory. *Signs: Journal of Women in Culture and Society*, *13*(3), 405–436. https://doi.org/10.1086/494426

Alcoff, L. (1991). The problem of speaking for others. *Cultural Critique*, *20*(1991–1992), 5–32. http

s://doi.org/10.2307/1354221

Ananny, M., & Gillespie, T. (2016). *Public platforms: Beyond the cycle of shocks and exceptions*. Oxford Internet Institute, The University of Oxford: The Internet, Policy & Politics Conferences.

Ashley, W. (2014). The angry Black woman: The Impact of pejorative stereotypes on psychotherapy with Black women. *Social Work in Public Health*, *29*(1), 27–34. https://doi.org/10.1080/19371918.2011.619449

Bakhtin, M. M. (1982). *The Dialogic Imagination: Four Essays: 1* (M. Holquist & C. Emerson, Eds.; Revised ed.). University of Texas Press. ISBN.

Bartolo, L. (2021). 'Eyes wide open to the context of content': Reimagining the hate speech policies of social media platforms through a substantive equality lens. *Renewal*, *29*(2), 39–51.

Bartolo, L., & Matamoros-Fernández, A. (forthcoming). *Online harms* (Platform Governance Terminologies). Information Society Project, Yale Law School. https://law.yale.edu/isp/publications/platform-governance-terminologies

BBC News. (2020). TikTok Holocaust trend "hurtful and offensive". *BBC News*. https://www.bbc.com/news/newsbeat-53934500

Beard, M. (2014). *Laughter in ancient Rome: On joking, tickling, and cracking up* (1st edition). University of California Press. https://www.jstor.org/stable/10.1525/j.ctt6wqbdt

Bell, M. C. (2021). John Stuart Mill's harm principle and free speech: Expanding the notion of harm. *Utilitas*, *33*(2), 162–179. https://doi.org/10.1017/S0953820820000229

Benesch, S. (2020). Proposals for improved regulation of harmful content online. In Y. Shany (Ed.), *Reducing online hate speech: Recommendations for social media companies and internet intermediaries* (pp. 247–306). Israel Democracy Institute.

boyd, danah. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (Ed.), *Networked self: Identity, community, and culture on social network sites* (pp. 39–58).

Brown, A. (2015). *Hate speech law: A philosophical examination*. https://doi.org/10.4324/9781315714899

Brown, A. (2017). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, *36*(4), 419–468. https://doi.org/10.1007/s10982-017-9297-1

Bunting, M. (2018). From editorial obligation to procedural accountability: Policy approaches to online content in the era of information intermediaries. *Journal of Cyber Policy*, *3*(2), 165–186. https://doi.org/10.1080/23738871.2018.1519030

Carlson, B., & Frazer, R. (2021). Fun. In B. Carlson & R. Frazer, *Indigenous digital life: The practice and politics of being Indigenous on social media* (pp. 121–139). Springer International Publishing. https://doi.org/10.1007/978-3-030-84796-8_6

Collins, P. H. (1990). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. Unwin Hyman.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against Women of Color. *Stanford Law Review*, *43*(6), 1241–1299. https://doi.org/10.2307/1229039

Critchley, S. (2002). *On humour*. Routledge. https://doi.org/10.4324/9780203870129

Culliford, E. (2020, March 5). Twitter bans posts that 'dehumanize' people in connection with diseases. *Reuters*. https://www.reuters.com/article/us-twitter-content-rule/twitter-bans-posts-that-dehumanize-people-in-connection-with-diseases-idUSKBN20S2K3

Davies, H., & Ilott, S. (Eds.). (2018). *Comedy and the politics of representation: Mocking the weak*. Springer International Publishing. https://doi.org/10.1007/978-3-319-90506-8

de Silva, A. (2020). Addressing the vilification of women: A functional theory of harm and implications for law. *Melbourne University Law Review*, *43*(3), 987–1032.

DeCook, J. R., Cotter, K., Kanthawala, S., & Foyle, K. (2022). Safe from "harm": The governance of violence by platforms. *Policy & Internet*, *14*(1), 63–78. https://doi.org/10.1002/poi3.290

Department for Digital, Culture, Media & Sport. (2022, November 28). *New protections for children and free speech added to internet laws* [Press release]. Online Safety, Society and Culture. https://www.gov.uk/government/news/new-protections-for-children-and-free-speech-added-to-internet-laws

Dias Oliva, T., Antonialli, D. M., & Gomes, A. (2021). Fighting hate speech, silencing drag queens? Artificial Intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, *25*(2), 700–732. https://doi.org/10.1007/s12119-020-09790-w

Digital, Culture, Media and Sport Committee. (2022). *The Draft Online Safety Bill and the legal but harmful debate (eighth report of session 2021–22)* (Report No. 8; Session 2021-22). UK Parliament. https://publications.parliament.uk/pa/cm5802/cmselect/cmcumeds/1039/report.html

Digital Industry Group Inc. (2021). *Australian Code of Practice on Disinformation and Misinformation*. Digital Industry Group Inc. https://digi.org.au/wp-content/uploads/2021/10/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL-WORD-UPDATED-OCTOBER-11-2021.pdf

Duguay, S., Burgess, J., & Suzor, N. (2020). Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence*, *26*(2), 237–252. https://doi.org/10.1177/1354856518781530

Dwoskin, E., Tiku, N., & Kelly, H. (2020, December 3). Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show. *The Washington Post*. https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/?utm_medium=social&utm_campaign=wp_main&utm_source=twitter

Edwards, L. (2021, October 11). Can the Online Safety Bill be more than a toothless tiger (or a Facebook flop)? [Blog post]. *Media@LSE*. https://blogs.lse.ac.uk/medialse/2021/10/11/can-the-online-safety-bill-be-more-than-a-toothless-tiger-or-a-facebook-flop/

Ellis, B. (1991). The last thing… Said: The challenger disaster jokes and closure. *International Folklore Review*, *8*, 110–124.

Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, no. COM/2020/825 final (2020). https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN

European Parliament. (2022, January 20). *Digital Services Act: Regulating platforms for a safer online space for users*[Press release]. News, European Parliament. https://www.europarl.europa.eu/news/en/press-room/20220114IPR21017/digital-services-act-regulating-platforms-for-a-safer-online-space-for-users

Facebook safety. (2013). *Controversial, harmful and hateful speech on Facebook*. Facebook. https://ww

w.facebook.com/notes/344897590097197/

Fave, L. (1977). Ethnic humour: From paradoxes towards principles. In A. Chapman & H. Foot (Eds.), *It's a funny thing, humor. Proceedings of The International Conference on Humour and Laughter 1976* (pp. 237–260). Pergamon Press. https://doi.org/10.1016/B978-0-08-021376-7.50049-8

Feinberg, J. (1987). *The moral limits of the criminal law. Volume 1: Harm to others* (1st ed.). Oxford University PressNew York. https://doi.org/10.1093/0195046641.001.0001

Feinberg, J. (1988). *The moral limits of the criminal law. Volume 2: Offense to others* (1st ed.). Oxford University PressNew York. https://doi.org/10.1093/0195052153.001.0001

Féret v. Belgium, Application no. 15615/07 (European Court of Human Rights 2009). https://hudoc.echr.coe.int/eng-press?i=003-2800730-3069797

Fielitz, M., & Ahmed, R. (2021). *It's not funny anymore. Far-right extremists' use of humour* [Report]. European Commission; Radicalisation Awareness Network; Publications Office of the European Union. https://home-affairs.ec.europa.eu/system/files/2021-03/ran_ad-hoc_pap_fre_humor_20210215_en.pdf

Ford, T. E., Boxer, C. F., Armstrong, J., & Edel, J. R. (2008). More than "just a joke": The prejudice-releasing function of sexist humor. *Personality and Social Psychology Bulletin*, *34*(2), 159–170. https://doi.org/10.1177/0146167207310022

Freeman, L., & Schroer, J. W. (Eds.). (2020). *Microaggressions and philosophy*. Routledge. https://doi.org/10.4324/9780429022470

Friedlaender, C. (2018). On microaggressions: Cumulative harm and individual responsibility. *Hypatia*, *33*(1), 5–21. https://doi.org/10.1111/hypa.12390

Fry, W. (1977). The appeasement function in mirthful laughter. In A. Chapman & H. Foot (Eds.), *It's a funny thing, humor. Proceedings of The International Conference on Humour and Laughter 1976* (pp. 23–26). Pergamon Press. https://doi.org/10.1016/B978-0-08-021376-7.50009-7

Gardner, J. (1989). Liberals and unlawful discrimination. *Oxford Journal of Legal Studies*, *9*(1), 1–22. https://doi.org/10.1093/ojls/9.1.1

Gatwiri, K. (2018, September 11). The cartoon of Serena Williams is the latest in a long line of harmful caricatures of Black women. *SBS*. https://www.sbs.com.au/topics/voices/culture/article/2018/09/11/cartoon-serena-williams-latest-long-line-harmful-caricatures-black-women

Gelber, K. (2021). Differentiating hate speech: A systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, *24*(4), 393–414. https://doi.org/10.1080/13698230.2019.1576006

Gelber, K., & McNamara, L. J. (2016). Anti-Vilification laws and public racism in Australia: Mapping the gaps between the harms occasioned and the remedies provided. *UNSW Law Journal*, *39*(2). https://www.unswlawjournal.unsw.edu.au/article/anti-vilification-laws-and-public-racism-in-australia-mapping-the-gaps-between-the-harms-occasioned-and-the-remedies-provided

George, A. (2022, February 28). *Insights: The Online Safety Bill* [News post]. News & Events. https://www.oii.ox.ac.uk/news-events/news/insights-the-online-safety-bill/

Ghaffary, S. (2020, August 11). Facebook bans Blackface and certain Anti-Semitic conspiracy theories. *Vox*. https://www.vox.com/recode/2020/8/11/21363815/facebook-bans-blackface-anti-semitic-conspiracy-theories

Gillespie, T. (2018a). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2018b). Regulation of and by platforms. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The SAGE handbook of social media* (pp. 254–278). SAGE Publications. https://doi.org/10.4135/9781473984066.n15

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 205395172094323. https://doi.org/10.1177/2053951720943234

Gillett, R. (2021). "This is not a nice safe space": Investigating women's safety work on Tinder. *Feminist Media Studies*, 1–17. https://doi.org/10.1080/14680777.2021.1948884

Godioli, A. (2020). Cartoon controversies at the European Court of Human Rights: Towards forensic humor studies. *Open Library of Humanities*, *6*(1), 22. https://doi.org/10.16995/olh.571

Godioli, A., & Little, L. E. (2022). Different systems, similar challenges: Humor and free speech in the United States and Europe. *HUMOR*, *35*(3), 305–327. https://doi.org/10.1515/humor-2021-0121

Goldman, E. (2021). Content moderation remedies. *Michigan Technology Law Review*, *28*(1), 1–59. https://doi.org/10.36645/mtlr.28.1.content

Gover, A. R., Harper, S. B., & Langton, L. (2020). Anti-Asian hate crime during the covid-19 pandemic: Exploring the reproduction of inequality. *American Journal of Criminal Justice*, *45*(4), 647–667. https://doi.org/10.1007/s12103-020-09545-1

Gray, C., & Hansen, K. (2021). Did covid-19 lead to an increase in hate crimes toward Chinese people in London? *Journal of Contemporary Criminal Justice*, *37*(4), 569–588. https://doi.org/10.1177/10439862211027994

Greene, V. S. (2019). "Deplorable" satire: Alt-Right memes, White genocide Tweets, and redpilling normies. *Studies in American Humor*, *5*(1), 31–69. https://doi.org/10.5325/studamerhumor.5.1.0031

Hage, G. (2014). Continuity and change in Australian racism. *Journal of Intercultural Studies*, *35*(3), 232–237. https://doi.org/10.1080/07256868.2014.899948

Hallinan, B. (2021). Civilizing infrastructure. *Cultural Studies*, *35*(4–5), 707–727. https://doi.org/10.1080/09502386.2021.1895245

Handyside v UK, Application no. 5493/72 (European Court of Human Rights 1976). https://hudoc.echr.coe.int/eng?i=001-57499

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575. https://doi.org/10.2307/3178066

Harding, S. (1992). Rethinking standpoint epistemology: What is strong objectivity? *The Centennial Review*, *36*(3), 437–470.

Hargrave, A. M., & Livingstone, S. (2006). Harm and offence in media content. *Intellect Ltd*.

Heinze, E. (2013). Hate speech and the normative foundations of regulation. *International Journal of Law in Context*, *9*(4), 590–617. https://doi.org/10.1017/S1744552313000311

Hope Not Hate. (2021). *Free speech for all: Why legal but harmful content should continue to be included in the Online Safety Bill* [Report]. Hope Not Hate. https://hopenothate.org.uk/wp-content/uploads/2021/09/Free-Speech-For-All-2021-08-v21Oct.pdf

IFRC, Unicef, & W.H.O. (2020). *Social stigma associated with COVID-19* [Guide]. UNICEF; World Health Organization; International Federation of Red Cross and Red Crescent Societies. https://www.who.int/publications/i/item/social-stigma-associated-with-covid-19

Index on Censorship. (2021, June 23). *Government's online safety bill will be "catastrophic for ordinary people's freedom of speech" says David Davis MP* [Press release]. https://www.indexoncensorship.org/2021/06/governments-online-safety-bill-will-be-catastrophic-for-ordinary-peoples-freedom-of-speech-says-david-davis-mp/

Joint Committee on the Draft Online Safety Bill. (2021). *Draft Online Safety Bill: Report of session 2021–22* (Paper HL 129; HC 609). UK Parliament. https://committees.parliament.uk/publications/8206/documents/84092/default/

Judson, E. (2022). *The Online Safety Bill: Demos position paper* [Position paper]. Demos. https://demos.co.uk/project/the-online-safety-bill-demos-position-paper/

Kahn, R. A. (2013). Why do Europeans ban hate speech: Debate between Karl Loewenstein and Robert Post. *Hofstra Law Review*, *41*(3), 545–586.

Kamp, A., Denson, N., Atie, R., Dunn, K., Sharples, R., Vergani, M., Walton, J., & Sisko, S. (2021). *Asian Australians' experiences of racism during the COVID-19 pandemic*. Centre for Resilient and Inclusive Societies. https://static1.squarespace.com/static/5d48cb4d61091100011eded9/t/6179e44a59f9670c25372139/1635378255065/Asian+Australians%27+Experiences+of+Racism+During+the+Covid-19+Pandemic.pdf

Keil, R., & Ali, H. (2006). Multiculturalism, racism and infectious disease in the global city: The experience of the 2003 sars outbreak in Toronto. *TOPIA: Canadian Journal of Cultural Studies*, *16*, 23–49. https://doi.org/10.3138/topia.16.23

Keller, D. (2015, October 12). Empirical evidence of "over-removal" by internet companies under intermediary liability laws [Blog post]. *The Center for Internet and Society, Stanford Law School*. http://cyberlaw.stanford.edu/blog/2015/10/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws

Kreuz, R. J., & Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, *118*(4), 374–386. https://doi.org/10.1037/0096-3445.118.4.374

Kuipers, G. (2011). The politics of humour in the public sphere: Cartoons, power and modernity in the first transnational humour scandal. *European Journal of Cultural Studies*, *14*(1), 63–80. https://doi.org/10.1177/1367549410370072

Larimore, S., Kennedy, I., Haskett, B., & Arseniev-Koehler, A. (2021). Reconsidering annotator disagreement about racist language: Noise or signal? *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, 81–90. https://doi.org/10.18653/v1/2021.socialnlp-1.7

Larsen, S. E. (2014). Towards the blasphemous self: Constructing societal identity in Danish debates on the Blasphemy Provision in the twentieth and twenty-first centuries. *Journal of Ethnic and Migration Studies*, *40*(2), 194–211. https://doi.org/10.1080/1369183X.2013.851471

Leerssen, P. (2021). Content curation. In L. Belli, N. Zingales, & Y. Curzi (Eds.), *Glossary of platform law and policy terms*. FGV Direito Rio; Internet Governance Forum. Glossary of Platform Law and Policy Terms

Little, L. E. (2008). Regulating funny: Humor and the law. *Cornell Law Review*, *94*(5), 1235–1292.

Lockyer, S., & Pickering, M. (2005). *Beyond a joke: The limits of humour*. Palgrave Macmillan UK.

Lott, E. (1992). Love and theft: The racial unconscious of Blackface minstrelsy. *Representations*, *39*, 23–50. https://doi.org/10.2307/2928593

Mallapragada, M. (2021). Asian Americans as racial contagion. *Cultural Studies*, *35*(2–3), 279–290. https://doi.org/10.1080/09502386.2021.1905678

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, *20*(6), 930–946. https://doi.org/10.1080/1369118X.2017.1293130

Matamoros-Fernández, A., Rodriguez, A., & Wikström, P. (2022). Humor that harms? Examining racist audio-visual memetic media on TikTok during covid-19. *Media and Communication*, *10*(2), 180–191. https://doi.org/10.17645/mac.v10i2.5154

Matsuda, M. J., Lawrence, C. R., Delgado, R., & Williams Crenshaw, K. (Eds.). (1993). *Words that wound: Critical race theory, assaultive speech, and the First Amendment*. Westview Press.

McClure, E. (2020). Escalating linguistic violence: From microaggressions to hate speech. In L. Freeman & J. W. Schroer (Eds.), *Microaggressions and Philosophy* (pp. 121–145). Routledge.

McGowan, M. K. (2009). Oppressive speech. *Australasian Journal of Philosophy*, *87*(3), 389–407. https://doi.org/10.1080/00048400802370334

McTernan, E. (2018). Microaggressions, equality, and social practices. *Journal of Political Philosophy*, *26*(3), 261–281. https://doi.org/10.1111/jopp.12150

Mendes, A. C. (2021). From "crisis" to imagination: Putting White heroes under erasure Post-George Floyd. *Cultural Studies ↔ Critical Methodologies*, *21*(5), 394–400. https://doi.org/10.1177/15327086211028677

Meta. (2022). *Facebook community standards, Hate speech*. Transparency Center. https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & Society*, *35*, 957–967. https://doi.org/10.1007/s00146-020-00950-y

Mill, J. S. (1978). *On liberty*. Open Road Integrated Media, Inc.

Mittelstadt, B. (2017). From individual to group privacy in big data analytics. *Philosophy & Technology*, *30*(4), 475–494. https://doi.org/10.1007/s13347-017-0253-7

Morreall, J. (1986). *The philosophy of laughter and humor*. State University of New York Press.

Morreall, J. (2009). *Comic relief: A comprehensive philosophy of humor*. https://doi.org/10.1002/9781444307795

Munn, L. (2020). Angry by design: Toxic communication and technical architectures. *Humanities and Social Sciences Communications*, *7*(1), 53. https://doi.org/10.1057/s41599-020-00550-7

Nagle, A. (2017). *Kill all normies*. Zero Books.

Nash, V. (2019a). Revise and resubmit? Reviewing the 2019 Online Harms White Paper. *Journal of Media Law*, *11*(1), 18–27. https://doi.org/10.1080/17577632.2019.1666475

Nash, V. (2019b, October 10). *Where's the harm in online hate speech?* SELMA. Hacking Hate. http

s://hackinghate.eu/news/where-s-the-harm-in-online-hate-speech/

Ndiaye, A. (2021, May 10). Together against covid-19 misinformation: A new campaign in collaboration with the WHO [Blog post]. *Meta for Media*. https://www.facebook.com/formedia/blog/together-against-covid-19-misinformation-a-new-campaign-in-partnership-with-the-who

Newton, C. (2020, March 5). Tech companies are getting more aggressive to fight COVID-19 hoaxes. *The Verge*. https://www.theverge.com/interface/2020/3/5/21164683/covid-19-tech-response-facebook-google-twitter-microsoft-youtube-whatsapp

Odijie, M. (2018). The fear of 'Yellow Peril' and the emergence of European Federalist Movement. *The International History Review*, *40*(2), 358–375. https://doi.org/10.1080/07075332.2017.1329751

Ono, K. A., & Pham, V. N. (2009). *Asian Americans and the media*. Polity.

O'Reilly, A. (2016). In defence of offence: Freedom of expression, offensive speech, and the approach of the European Court of Human Rights. *Trinity College Law Review*, *19*, 234–260.

Otto-Preminger-Institut v. Austria, Application no. 13470/87 (European Court of Human Rights 1994). https://hudoc.echr.coe.int/eng?i=001-57897

Oversight Board. (2021, May 20). *Case decision 2021-005-FB-UA*. News and Articles. https://oversightboard.com/news/923336265153853-oversight-board-overturns-facebook-decision-case-2021-005-fb-ua/

Oversight Board. (2022, June 18). *Case decision 2022-001-FB-UA*. https://oversightboard.com/news/1629549600777906-oversight-board-overturns-meta-s-original-decision-in-knin-cartoon-case-2022-001-fb-ua/

Paasonen, S., Jarrett, K., & Light, B. (2019). *NSFW: Sex, humor, and risk in social media*. The MIT Press. https://doi.org/10.7551/mitpress/10916.001.0001

Pemberton, S. A. (2015). *Harmful societies: Understanding social harm*. Policy Press. https://doi.org/10.1332/policypress/9781847427946.001.0001

Phillips, W., & Milner, R. M. (2017). *The ambivalent internet: Mischief, oddity, and antagonism online* (1st ed.). Polity.

Ramsey, F. (2020, June 28). *Hey @youtube why have some of the biggest creators on the platform been allowed to monetize videos with "jokes" about sexual abuse, pedophilia, racial slurs & blackface? And not one video. HUNDREDS of videos with MILLIONS of views? What was the ad revenue share on your end?* [Tweet]. @chescaleigh. https://twitter.com/chescaleigh/status/1277147552199192576

Renwick, A. (2021). *Dr Renwick gives evidence to the DCMS Sub-committee on Online Harms and Disinformation*. The UCL Constitution Unit. https://www.ucl.ac.uk/constitution-unit/news/2021/sep/dr-renwick-gives-evidence-dcms-sub-committee-online-harms-and-disinformation

Riley, J. (2009). *Racism, blasphemy, and free speech* (C. L. Ten, Ed.). Cambridge University Press.

Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. In S. U. Noble & B. M. Tynes (Eds.), *The intersectional internet: Race, sex, class, and culture online* (pp. 147–159). Peter Lang Publishing.

Rostbøll, C. F. (2008, August 26). *The use and abuse of 'universal values' in the Danish cartoon controversy*. Conference 'Democracy and Difference', Culcom.

Schmon, C. (2021, July 14). *UK's draft online safety bill raises serious concerns around freedom of*

expression.

Schoenebeck, S., Haimson, O. L., & Nakamura, L. (2021). Drawing from justice theories to support targets of online harassment. *New Media & Society*, *23*(5), 1278–1300. https://doi.org/10.1177/1461444820913122

Shifman, L. (2014). *Memes in digital culture*. The MIT Press.

Siapera, E., & Viejo-Otero, P. (2021). Governing hate: Facebook and digital racism. *Television & New Media*, *22*(2), 112–130. https://doi.org/10.1177/1527476420982232

Smith, S. D. (2006). Is the harm principle illiberal. *American Journal of Jurisprudence*, *51*(1), 1–42. https://doi.org/10.1093/ajj/51.1.1

Smuha, N. A. (2021). Beyond the individual: Governing AI's societal harm. *Internet Policy Review*, *10*(3). https://doi.org/10.14763/2021.3.1574

Stevens, A. (2007). Carnival and comedy: On Bakhtin's Misreading of Boccaccio. *Opticon1826*, *3*. https://doi.org/10.5334/opt.030707

Stott, A. (2004). *Comedy* (1st ed.). Routledge. https://doi.org/10.4324/9780203312124

Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge University Press. https://doi.org/10.1017/9781108666428

Suzor, N. P., Myers West, S., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, 1526–1543. https://doi.org/1932–8036/20190005

Thampapillai, D. (2010). *Low level racist speech: Beyond law's reach?* (No. 1551567). SSRN. https://doi.org/10.2139/ssrn.1551567

Thomae, M., & Viki, G. T. (2013). Why did the woman cross the road? The effect of sexist humor on men's rape proclivity. *Journal of Social, Evolutionary, and Cultural Psychology*, *7*(3), 250–269. https://doi.org/10.1037/h0099198

Turillazzi, A., Casolari, F., Taddeo, M., & Floridi, L. (2022). *The Digital Services Act: An analysis of its ethical, legal, and social implications* (No. 4007389). SSRN. https://doi.org/10.2139/ssrn.4007389

Vidgen, B., Burden, E., & Margetts, H. (2021). *Understanding online hate*. The Alan Turing Institute. https://www.ofcom.org.uk/__data/assets/pdf_file/0022/216490/alan-turing-institute-report-understanding-online-hate.pdf

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Weaver, S. (2011). Liquid racism and the ambiguity of Ali G. *European Journal of Cultural Studies*, *14*(3), 249–264. https://doi.org/10.1177/1367549410396004

Wolfe, P. (2016). *Traces of history: Elementary structures of race*. Verso.

Wolsey, C. (2020). *My (racism) apology: A case study of YouTubers' apology videos for racist acts* [Undergraduate research paper, MacEwan]. Ro@m. https://hdl.handle.net/20.500.14078/42

Appendix

To map platforms' policies related to humour, satire and parody, we ran keyword searches of their user guidelines. In Table 1, we list the keywords used, the URL to each of the platforms' guidelines and the dates the searches were run.

TABLE 1: The keyword search underlying our policy mapping exercise

KEYWORDS	[HUMOR] [HUMOUR] [JOKE] [SATIRE] [SATIRICAL] [IRONY] [IRONIC] [MOCK] [LAUGH] [FUNNY] [MAKE FUN] [MAKES FUN] [TEASE] [TEASING] [PARODY] [MEME] [COMEDY]			
SEARCH DATE	18 AUGUST 2022			
PLATFORM	Facebook	Twitter	YouTube	TikTok
USER GUIDELINES	https://www.facebook.com/communitystandards/	https://help.twitter.com/en/rules-and-policies/twitter-rules	https://www.youtube.com/intl/ALL_au/howyoutubeworks/policies/community-guidelines/	https://www.tiktok.com/community-guidelines

Tables 2-5 map each of the platforms' policies in relation to humour, satire and parody. Based on the results of our keyword searches, we grouped policies into three categories:

- A general 'Humour' category, which captures policies that refer to humorous expression, but also 'mocking' and 'laughing', for example
- A 'Satire' category, which only captures policies that specifically refer to satire/ satirical expression
- A 'Parody' category, which only includes policies that specifically refer to 'parody'

TABLE 2: Facebook's Community Standards

	FACEBOOK		
	POLICY	*MENTIONED AS EXCEPTION?*	*EXCERPTS FROM FACEBOOK'S COMMUNITY STANDARDS*
H	Suicide and	*N*	*"We also remove content that identifies and negatively targets victims or survivors of*

	FACEBOOK		
	POLICY	*MENTIONED AS EXCEPTION?*	*EXCERPTS FROM FACEBOOK'S COMMUNITY STANDARDS*
U M O U R	Self-Injury		*suicide or self-injury seriously, **humorously** or rhetorically"*
		N	*"We remove any content that encourages suicide or self-injury, including fictional content such as **memes** or illustrations and any self-injury content that is graphic, regardless of context."*
	Child Sexual Exploitation, Abuse and Nudity	*N*	*"Do not post: - Content that identifies or **mocks** alleged victims of child sexual exploitation by name or image"*
	Adult Sexual Exploitation	*N*	*"Do not post: - Content "**Mocking** victims of... non-consensual sexual touching, crushing, necrophilia or bestiality, or forced stripping" - "Secretly taken non-commercial imagery of a real person's commonly sexualized body parts (breasts, groin, buttocks, or thighs) or of a real person engaged in sexual activity. This imagery is commonly known as "creepshots" or "upskirts" and includes photos or videos that **mocks**, sexualizes or exposes the person depicted in the imagery" "For the following content, we include a warning screen so that people are aware the content may be disturbing: - Content **mocking** the concept of non-consensual sexual touching"*
	Bullying and Harassment	*N*	*"Do not: Target public figures by purposefully exposing them to: - Content that praises, celebrates or **mocks** their death or medical condition. Target private individuals or limited scope public figures with: - Content that praises, celebrates, or **mocks** their death or serious physical injury" "Do not: - Post content praising, celebrating or **mocking** anyone's death."*
	Hate Speech	*N*	*"Do not post: Content "**Mocking** the concept, events or victims of hate crimes even if no real person is depicted in an image."*
		N	*"Do not post: Content targeting a person or group of people on the basis of their protected characteristic(s) with claims that they have or spread the novel coronavirus, are responsible for the existence of the novel coronavirus, are deliberately spreading the novel coronavirus or **mocking** them for having or experiencing the novel coronavirus."*
		Y	*"In certain cases, we will allow content that may otherwise violate the Community Standards when it is determined that the content is **satirical**. Content will only be allowed if the violating elements of the content are being **satirized** or attributed to something or someone else in order to **mock** or criticize them."*
	Adult Nudity and Sexual Activity	*Y*	*"... we default to removing sexual imagery to prevent the sharing of non-consensual or underage content. Restrictions on the display of sexual activity also apply to digitally created content unless it is posted for educational, **humorous**, or satirical purposes."*
	Sexual solicitation	*Y*	*"Do not post: Sexually explicit language that goes into graphic detail beyond mere reference to: • A state of sexual arousal (e.g wetness or erection) or • An act of sexual intercourse (e.g sexual penetration, self-pleasuring or exercising fetish scenarios). Except for content shared in a **humorous**, satirical or educational context, as a sexual metaphor or as sexual cursing."*
S	Fraud and	*Y*	*"In certain cases, we will allow content that may otherwise violate the Community*

	FACEBOOK		
	POLICY	*MENTIONED AS EXCEPTION?*	*EXCERPTS FROM FACEBOOK'S COMMUNITY STANDARDS*
ATIRE	deception		*Standards when it is determined that the content is **satirical**. Content will only be allowed if the violating elements of the content are being **satirised** or attributed to something or someone else in order to **mock** or criticise them."*
	Restricted goods and services	*Y*	*"In certain cases, we will allow content that may otherwise violate the Community Standards when it is determined that the content is **satirical**. Content will only be allowed if the violating elements of the content are being **satirised** or attributed to something or someone else in order to **mock** or criticise them."*
	Dangerous individuals and organisations	*Y*	*"In certain cases, we will allow content that may otherwise violate the Community Standards when it is determined that the content is **satirical**. Content will only be allowed if the violating elements of the content are being **satirised** or attributed to something or someone else in order to **mock** or criticise them."*
	Privacy violations	*Y*	*"In certain cases, we will allow content that may otherwise violate the Community Standards when it is determined that the content is **satirical**. Content will only be allowed if the violating elements of the content are being **satirised** or attributed to something or someone else in order to **mock** or criticise them."*
	Sexual solicitation	*Y*	*"In certain cases, we will allow content that may otherwise violate the Community Standards when it is determined that the content is **satirical**. Content will only be allowed if the violating elements of the content are being **satirised** or attributed to something or someone else in order to **mock** or criticise them."*
	Adult Nudity and Sexual Activity	*Y*	*"... we default to removing sexual imagery to prevent the sharing of non-consensual or underage content. Restrictions on the display of sexual activity also apply to digitally created content unless it is posted for educational, humorous, or **satirical** purposes."*[25]

TABLE 3: Twitter's Rules

	TWITTER		
	POLICY	*MENTIONED AS EXCEPTION?*	*EXCERPTS FROM TWITTER RULES*
HUMOUR	Civic integrity policy	*N*	*"Given the significant risks of confusion about key election information, we may take these actions [removal or adding label] even if Tweets contain (or attempt to contain) **satirical** or **humorous** elements."*
PARODY	Misleading and deceptive identities policy	*Y*	*"The following, for example, are not in violation of this policy:... **Parody**, commentary, or fan accounts that comply with our policy for such accounts."* *"Accounts are less likely to violate this policy if the profile contains context that indicates the account is not affiliated with the subject in the profile image, as with **parody**, commentary, or fan accounts."*
	Civic integrity policy (false/ misleading affiliation)	*N*	*"You can't create fake accounts which misrepresent their affiliation, or share content that falsely represents its affiliation, to a candidate, elected official, political party, electoral authority, or government entity. Read more about our **parody**, commentary, and fan account policy."*
	Civic integrity	*Y*	*"Not all false or untrue information about politics or civic processes constitutes*

25. Note that this content is only shown to "individuals aged 18 and over".

	TWITTER		
	POLICY	*MENTIONED AS EXCEPTION?*	*EXCERPTS FROM TWITTER RULES*
	policy (false/ untrue information)		*manipulation or interference. In the absence of other policy violations, the following are generally not in violation of this policy:* *- using Twitter pseudonymously or as a **parody**, commentary, or fan account to discuss elections or politics."*
	Trademark policy	*Y*	*"Referencing another's trademark is not automatically a violation of Twitter's trademark policy. Examples of non-violations include:* *-using a trademark in a nominative or other fair use manner. For more information, see our **parody**, newsfeed, commentary, and fan account policy."*
	Copyright policy	*Y*	*"If you are concerned about the use of your brand or entity's name, please review Twitter's trademark policy. If you are concerned about a **parody**, newsfeed, commentary, or fan account, please see the relevant policy here. These are generally not copyright issues."*
SATIRE	Synthetic and manipulated media policy	*Y*	*"In the absence of other policy violations, the following are generally not in violation of this policy:... Memes or **satire**, provided these do not cause significant confusion about the authenticity of the media"*

TABLE 4: YouTube's Community Guidelines[26]

	YOUTUBE		
	POLICY	*MENTIONED AS EXCEPTION?*	*EXCERPTS FROM YOUTUBE'S COMMUNITY GUIDELINES*
HUMOUR	Guidelines for third-party assessors to determine content 'Quality'	*N*	*"For some topics, such as **humor** or recipes, less formal expertise is OK. For these topics, popularity, user engagement, and user reviews can be considered evidence of reputation. For topics that need less formal expertise, websites can be considered to have a positive reputation if they are highly popular and well-loved for their topic or content type, and are focused on helping users." (p. 30)*
SATIRE	Harassment and cyberbullying	*Y*	*"If the primary purpose is educational, documentary, scientific, or artistic in nature, we may allow content that includes harassment. These exceptions are not a free pass to harass someone. Some examples include:* *-- Scripted performances: Insults made in the context of an artistic medium such as scripted **satire**, stand up comedy, or music (such as a diss track). Note: This exception is not a free pass to harass someone and claim "I was joking."*
	Misinformation	*Y*	*"We may also make exceptions if the purpose of the content is to condemn, dispute, or **satirize** misinformation that violates our policies."*
	Elections misinformation	*Y*	*"We may allow content that violates the election integrity policy noted on this page if the content includes additional context... Additional context may include countervailing views, or if the purpose of the content is to condemn, dispute, or **satirize** misinformation that violates our policies."*
	COVID-19	*Y*	*"We may also make exceptions if the purpose of the content is to condemn, dispute, or*

26. Note that the 'Guidelines for third-party assessors to determine content 'Quality'' are taken from Google's Public Guidelines which describe YouTube's approach to tackling misinformation by reducing 'low quality' content in search and recommendations.

	YOUTUBE		
	POLICY	*MENTIONED AS EXCEPTION?*	*EXCERPTS FROM YOUTUBE'S COMMUNITY GUIDELINES*
	medical misinformation policy		***satirize** misinformation that violates our policies."*

TABLE 5: TikTok's Community Guidelines

	TIKTOK		
	POLICY	*MENTIONED AS EXCEPTION?*	*EXCERPTS FROM TIKTOK'S COMMUNITY GUIDELINES*
HUMOUR	Bullying and harassment	*N*	*"We remove all expressions of abuse, including threats or degrading statements intended to **mock**, humiliate, embarrass, intimidate, or hurt an individual. This prohibition extends to the use of TikTok features. To enable expression about matters of public interest, critical comments of public figures may be allowed; however, serious abusive behavior against public figures is prohibited."*
SATIRE	Community guidelines (overall)	*Y*	*"...we may allow exceptions under certain circumstances, such as educational, documentary, scientific, or artistic content, **satirical** content, content in fictional settings, counterspeech, or content that otherwise enables individual expression on topics of social importance."*
PARODY	Impersonation	*Y*	*"We do allow accounts that are clearly parody, commentary, or fan-based, such as where the username indicates that it is a fan, commentary, or **parody** account and not affiliated with the subject of the account."*
	Copyright and trademark infringement	*Y*	*"The use of copyrighted work under certain circumstances, such as the fair use doctrine or other applicable laws, or the use of a trademark to reference, lawfully comment, criticize, **parody**, make a fan page, or review a product or service may not be considered a violation of our policies."*

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