Online Harm

Louisa Bartolo & Ariadna Matamoros-Fernandez

introduction

Social media platforms, like Facebook, Instagram and YouTube, moderate "harmful" user-generated content and online behavior that would be considered illegal in many parts of the world (e.g., calls to violence), as well as other types of content and behavior that are deemed harmful but are not illegal (e.g., self-harm content). This moderation process is notoriously opaque and has invited widespread distrust. Platforms have been repeatedly accused of over-removing content, of under-removing it, of removing the wrong things, and of underplaying the extent to which their own systems incentivize and reward harmful content and behavior.[1] As a result, different regulatory attempts by national governments (e.g., Australia, Canada, the UK, Ireland) and supranational entities (e.g., the European Commission) seek to hold platforms accountable for "harmful" user-generated content and online behavior through various forms of "notice-and-action," risk assessment, transparency, and auditing requirements, often under the threat of fines.[2]

Historically, establishing a "harm" threshold has been key to determining the boundaries of legitimate regulatory intervention, and yet the way regulatory bodies and platforms define "harm," much less "online harm," is not self-evident.[3] The stakes of this definitional debate became clear when some jurisdictions, notably the UK, indicated plans to adopt an expansive definition of "harm" in their online safety regulation to

[1] *See* Tarleton Gillespie, *Platforms Are Not Intermediaries*, 2 GEO. L. TECH. REV. 198 (2018).

[2] These regulatory attempts are at different stages.

[3] *See* Victoria Nash, *Revise and Resubmit? Reviewing the 2019 Online Harms White Paper*, 11 J. MEDIA L. 18, 22 (2019); Julia R. DeCook, Kelley Cotter, Shaheen Kanthawala & Kali Foyle, *Safe from "Harm": The Governance of Violence by Platforms*, 14 POL'Y & INTERNET 63 (2022).

cover not only illegal content and conduct that causes harm (e.g., terrorism or violence-inciting speech) but also "legal but harmful content," such as adult-directed abuse that falls below criminal thresholds and self-harm promotion.[4]

The inclusion of a "legal but harmful content" category within the UK's Online Safety Bill draft received extensive backlash for its potential to facilitate state censorship of legitimate and legally protected expression, and free speech proponents advocated scrapping it from the Bill.[5] A new draft of the Bill was reintroduced in parliament in January 2023.[6] In this amended version, duties relating to "legal but harmful" content accessed by adults were removed from the legislation: platforms will no longer be duty-bound to produce risk assessments relating to specific types of "legal but harmful" content and behavior on their services.[7] Some of what would have been considered "legal but harmful" material may become illegal following the government's announcement that they are looking to criminalize "the encouragement of self-harm and

[4] *See* Press Release, Nadine Dorries, Secretary of State, UK Dep't for Digit., Culture, Media & Sport, Statement Made on 7 July 2022 [Statement UIN HCWS194] (July 7, 2022), https://questions-statements.parliament.uk/written-statements/detail/2022-07-07/hcws194.

[5] For example, Index on Censorship is an organization campaigning to "[l]imit online regulation to addressing illegal content." Ruth Anderson, *#OffOn – Don't Switch Off Our Online Rights*, INDEX ON CENSORSHIP (Oct. 11, 2021), https://www.indexoncensorship.org/2021/10/offon-dont-switch-off-our-online-rights. The civil society group Big Brother Watch argued that, by including "legal but harmful" content within its scope, the Bill will "cause platforms to significantly expand their already-censorious content policies." *Online Safety Bill: What the Government Must Do Next*, BIG BROTHER WATCH (Feb. 10, 2022), https://bigbrotherwatch.org.uk/2022/02/the-online-safety-bill-what-the-government-must-do-next. Both organizations, along with other civil society groups such as Article 19 and Open Rights Group, wrote a letter to the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression in late 2022 in which they warned about the dangers of a vague "legal but harmful" category setting the stage for undue state censorship. *"The Online Safety Bill Will Fundamentally Undermine Rights to Freedom of Expression": Index and Other Organisations Ask UN Special Rapporteurs to Intervene on Proposed UK Legislation*, INDEX ON CENSORSHIP (Nov. 16, 2022), https://www.indexoncensorship.org/2022/11/the-online-safety-bill-will-fundamentally-undermine-rights-to-freedom-of-expression.

[6] Online Safety Bill 2022-23, HL Bill [87] (Rev) (U.K.)

[7] This represents an important shift from earlier versions of the Bill, which went beyond a focus on illegal harms and set the stage for there to be an evolving, parliament-approved list of lawful online harms set out in secondary legislation. In this earlier version of the Online Safety Bill, the largest (so-called "Category 1") platforms were being called on to produce risk assessments and transparently disclose (via their Terms of Service) how they would address this list of lawful harms.

the sharing of people's intimate images without their consent."[8] For other content categories that do not meet criminal thresholds and are no longer defined as "harmful" for the purposes of the Bill – "such as the glorification of eating disorders, racism, antisemitism or misogyny" – internet companies will need to offer more user controls to help people avoid seeing this content.[9] Platforms will also be expected to enforce their own Terms of Service—effectively leaving it to platforms to decide which, if any, lawful harms they address, and prioritizing a narrow focus on content moderation, to the exclusion of systemic and design-based approaches.[10] For reasons we set out in more detail below, we believe that dropping the "legal but harmful" provisions from the new Bill is a missed opportunity.

In this essay, we make the case for expansive conceptualizations of "online harm" in online safety regulation that go beyond dominant liberal legal frameworks, and that incentivize platforms to adopt a range of different and proportionate remedies to address lawful harms on their services.[11] In particular, we suggest that by working in meaningful consultation with civil society groups and platform companies, regulatory bodies[12] have an opportunity to conceptualize (lawful) online harms in a way that better reflects the needs of those most affected by these harms, deals with real risks of societal harm, and takes historically entrenched

[8] Press Release, UK Department for Digital, Culture, Media & Sport & the Rt Hon Michelle Donelan MP, New Protections for Children and Free Speech Added to Internet Laws (Nov. 28, 2022),

https://www.gov.uk/government/news/new-protections-for-children-and-free-speech-added-to-internet-laws.

[9] *Id.*

[10] *See* Lorna Woods, William Perrin & Maeve Walsh, *Online Safety Bill – Government Amendments for Committee Stage*, CARNEGIE UK TRUST (Dec. 5, 2022), https://www.carnegieuktrust.org.uk/blog-posts/online-safety-bill-indicative-amendments.

[11] *See* Eric Goldman, *Content Moderation Remedies*, 28 MICH. TECH. L. REV. 1 (2021).

[12] In this essay, we assume a democratic state. Nevertheless, we are alive to the fact that states, including democratic states, are in various instances perpetrators of harm towards citizens themselves. Concern about the state having undue influence over platform governance is therefore warranted. However, state-approved regulation comes in many forms. For example, the UK communications regulator Ofcom is independent of, but answerable to, the UK Parliament. If empowered to do so by the Online Safety Bill, Ofcom could mandate risk assessments and transparency requirements for platforms to incentivize responsible platform governance around "legal but harmful" content and behavior without prescribing specific remedies to be used by platforms.

power relations seriously.[13] We are alive to concerns that expansive conceptualizations of "harm" invite risks of regulatory overreach on the part of states but argue that retreating to narrow, legal conceptions of "(online) harm" in state regulation falls short of protecting users.

The UK Online Safety Bill draft and the intense debates it has spurred raise critical issues of relevance far beyond the UK. 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. Definitions of "harm" are neither pre-ordained nor static,[14] which invites more critical consideration of *who* is being given the power to define "harms" in emerging online safety regulation and platform policy, and whose perspectives may be being sidelined or delegitimized when certain conceptions of "harm,"[15] and certain evidentiary standards of "harm,"[16] are privileged over others. The UK's example also highlights a common misunderstanding around what passing legislation that pushes platforms to do better in their moderation of legal but harmful speech and conduct would entail. As many UK groups have already highlighted, their advocacy to maintain the "legal but harmful" category in the Online Safety Bill does not mean they want this content and behavior to receive the same treatment as illegal content.[17] Instead, a "systems and processes"

[13] *See* Sarina Schoenebeck, Oliver L. Haimson & Lisa Nakamura, *Drawing from Justice Theories to Support Targets of Online Harassment*, 23 NEW MEDIA & SOC'Y 1278 (2021). The question of what makes consultation meaningful is of course open to debate, but there are ways to make these processes more inclusive, transparent, and open to scrutiny. With regard to the role of the state in this process, states can produce a more, or less, enabling environment for civil society groups, and can take steps to ensure that more civil society players are brought to the table when important discussions about platform governance are taking place. Given the international scope of platforms, nation-states can be an important lever to ensure that civil society representation in platform governance is more geographically representative. For a nuanced perspective on the ways that states can enable (and hinder) meaningful civil society participation in platform governance, see Brenda Dvoskin, *Representation Without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance*, 67 VILL. L. REV. 447 (2022).

[14] *See* John Gardner, *Liberals and Unlawful Discrimination*, 9 OXFORD J. LEGAL STUD. 1 (1989); Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, & Jed R. Brubaker, *Understanding International Perceptions of the Severity of Harmful Content Online*, 16 PLoS ONE 1 (2021).

[15] *See* Steven D. Smith, *Is the Harm Principle Illiberal?*, 51 AM. J. JURIS. 1 (2006).

[16] *See* J. Nathan Matias, Eric Pennington & Zenobia T. Chan, *Testing Concerns about Technology's Behavioral Impacts with N-of-One Trials* (2022).

[17] we refer here to the work of civil society groups like demos, glitch and carnegie uk trust. for examples of where they have advocated for a systems-and-processes based approach to addressing "legal but harmful" content and behavior, rather than take-downs, see Ellen

regime, elements of which were originally put forward in the UK and which has been adopted in relation to "systemic risks" in the EU Digital Services Act (DSA), could require platforms to produce risk assessments[18] and to transparently disclose how they are addressing various forms of lawful harms, including those that are fueled by platforms' own technical affordances (e.g., via frictionless sharing or algorithmic amplification). Platforms are, in many cases, already addressing various forms of "legal but harmful" content and conduct, but they are doing so selectively and with little to no public oversight or accountability.

We structure the remainder of this paper as follows: First, we describe "harm" as an "essentially contested concept"[19] and discuss the long-recognized need for critical engagement with the underlying *partiality* and limitations of legal perspectives on "harm." Second, we discuss how, by embracing expansive conceptualizations of harm, online safety regulation could incentivize platforms to recognize, document, and address the risks of harm occurrence online more expansively.

I. The Significance of Calling Something a "Harm"

Historically, in liberal democratic societies, the "harm" threshold has been key to both justifying and delimiting the legitimate boundaries of regulatory intervention.[20] In nineteenth-century England, liberal theorist John Stuart Mill famously proposed "the harm principle" as the basis for

Judson, The Future of Legal but Harmful Remains Uncertain, Demos (Sep. 23, 2022), https://demos.co.uk/blog/what-we-can-expect-when-the-online-safety-bill-returns; and Why Legal But Harmful Content Should Continue to Be Included in the Online Safety Bill, Hope not Hate (Sep. 3, 2021), https://hopenothate.org.uk/2021/09/03/new-report-free-speech-for-all-why-legal-but-harmful-content-should-continue-to-be-included-in-the-online-safety-bill. See also Lorna Woods, William Perrin & Maeve Walsh, Submission to the Online Safety Bill Committee, Carnegie UK Trust (May 2022), https://www.carnegieuktrust.org.uk/publications/submission-to-online-safety-bill-committee.

[18] In the case of the UK, the updated Online Safety Bill looks set to retain platforms' risk assessment duties in relation to content and conduct deemed to pose a "danger" to children. It looks as though platforms will no longer be expected to produce risk assessments in relation to material, behavior and systems posing a risk to *adults.* Press Release, New Protections for Children and Free Speech Added to Internet Laws, *supra* note 8.

[19] W.B. Gallie, *Essentially Contested Concepts*, 56 PROC. ARISTOTELIAN SOC'Y 167 (1956).

[20] *See* Gardner, *supra* note 14; Victoria Nash, *Where's the Harm in Online Hate Speech?*, SELMA: HACKING HATE (Oct. 10, 2019), https://hackinghate.eu/news/where-s-the-harm-in-online-hate-speech.

legitimate societal and state intervention into individual affairs. Mill argued that a person's individual liberty could only be curtailed to prevent that individual from causing "harm" to others.[21] A century later, the American legal theorist Joel Feinberg would elaborate on Mill's principle in the context of criminal law, defining harm as the wrongful "thwarting, setting back, or defeating of an interest,"[22] especially individuals' "welfare interests." For Feinberg, "welfare interests" were the necessary conditions for individuals to sustain their version of a "good life," including "the absence of absorbing pain and suffering" (physical harm) and "the capacity to engage normally in social intercourse and to enjoy and maintain friendships."[23] Beyond criminal law, regulation related to everything from the environmental sector and workplace safety to the financial domain focuses on identifying and, to the extent possible, reducing the risk of "harms."[24] Applying the label "harm" in the context of regulatory debates is an inescapably political affair—much more than a description of an event or a person's experience, the term helps to establish what sorts of activities the state can legitimately involve itself in, discourage, and even prohibit.

In practice, relying on notions of "harm" to determine the legitimate bounds of regulation turns out to be tricky because there is little agreement about what counts as "harm." "Harm" is an "essentially contested concept"[25]—a term originally coined by W.B. Gallie to refer to concepts "the proper use of which inevitably involves endless disputes about their proper uses on the part of their users."[26] Joanne Conaghan notes that within the legal profession "much more effort" has been invested "into the business of deploying law as an instrument for the redress of harm than to more fundamental questions of what precisely harm entails and how we know and recognize its occurrence."[27] In their famous account of the stages involved in a legal dispute, Felstiner and colleagues noted that "the first stage of the disputing process—the

[21] JOHN STUART MILL, ON LIBERTY (1859).

[22] Joel Feinberg, *The Moral Limits of the Criminal Law: Harm to Others* 33 (1987).

[23] *Id.* at 37.

[24] Zohar Efroni, *The Digital Services Act: Risk-Based Regulation of Online Platforms*, INTERNET POL'Y REV. (Nov. 16, 2021), https://policyreview.info/articles/news/digital-services-act-risk-based-regulation-online-platforms/1606.

[25] *See* Jeremy Waldron, *Is the Rule of Law an Essentially Contested Concept (In Florida)?*, 21 L. & PHIL. 137 (2002).

[26] Gallie, *supra* note 19, at 169.

[27] Joanne Conaghan, *Law, Harm and Redress: A Feminist Perspective*, 22 LEGAL STUD. 319, 321 (2002).

perception of harm—was the least examined but perhaps the most important."[28] Ultimately, beliefs about what negative effects qualify as harms are contingent. Conceptions of harm are not static. Instead, ideas of "harm" evolve over time, not least as social movements help to shift social perceptions of what are acceptable or tolerable conditions to live a "good life." This advocacy sometimes results in legal reform, but not always.[29]

Acknowledging the term "harm" as a site of contestation means accepting that definitions of harm are never neutral: there is no singular "view from nowhere" when it comes to defining harm. Insights from feminist standpoint theory[30] are instructive in this regard, especially in the theory's insistence that the generation of knowledge (in this case what harm is or means) is always socially situated. Diverse and intersecting social positions (e.g., race, sexuality, religion, age), from which values are interpreted and constructed, impact people's ability to know about and define the world. For example, relying on existing legal notions of harm to define "online harms" ignores the ways in which legal systems have at various points been complicit in reproducing, rather than challenging, forms of oppression towards historically marginalized individuals and groups.[31] Power dynamics are critical to consider when evaluating competing discourses on harm and any process to draw definitional boundaries around the term requires a strong degree of reflexivity about who is being given the power to draw those boundaries, their social positions, and which voices those definitional boundaries might be excluding.

[28] Anna-Maria Marshall, *Confronting Sexual Harassment: The Law and Politics of Everyday Life* 23 (2005).

[29] *Id.*

[30] *See* Sandra Harding, *The Feminist Standpoint Theory Reader: Intellectual and Political Controversies* (2004); Patricia Hill Collins, *Learning from the Outsider Within: The Sociological Significance of Black Feminist Thought*, 33 SOC. PROBS. 14 (1986); Aileen Moreton-Robinson, *Towards an Australian Indigenous Women's Standpoint Theory*, 28 AUSTL. FEMINIST STUD. 331 (2013).

[31] *See* Melina Constantine Bell, *John Stuart Mill's Harm Principle and Free Speech: Expanding the Notion of Harm*, 33 UTILITAS 162 (2021).

2. Delimiting the Boundaries of "Online Harm”

Liberal notions of harm (including those espoused by Mill and Feinberg) have strong echoes in Anglo-American criminal law,[32] and often hinge on the distinction between something as “harmful” and a range of other experiences which, while unpleasant, hurtful, or undesirable, do not meet the threshold of being a “harm.” In the context of speech, which concerns much of the online safety regulation, Mill[33] and Feinberg[34] viewed liberty of expression as fundamental to individual flourishing, and they only recognized a harmful nature to speech acts in very limited circumstances: such as when speech constituted clear incitement to violence and defamation. According to this view, regulators would punish speech that incites imminent physical violence against someone on the basis of their race but allow “low level” racist jokes to circulate freely due to their “non-harmful” nature. Socio-legal scholars have, however, highlighted the limitations of this approach when it comes to addressing “routine” and “subtle” forms of abuse which do not meet legal harm thresholds but have “cumulative effects” on historically marginalized groups.[35]

Cumulative harm describes the case whereby an individual can be said to have suffered harm as a result of the repeated experience of several negative effects (which might not each by themselves be considered harmful by the standards of criminal law frameworks). The repetition of the negative experiences, and their “relational nature”—the fact that they “intensify one another in the process of accumulation”—is what makes them cumulatively harmful.[36] Microaggressions, a term coined in psychology used to describe “small acts of insult or indignity, relating to a person’s membership in a socially oppressed group” are often put forward as an example of behavior that may seem “minor on its own” but

[32] *SEE* Nina Persak, Criminalising Harmful Conduct: The Harm Principle, its Limits and Continental Counterparts (2007).

[33] *See* Jonathan Riley, *Racism, Blasphemy, and Free Speech*, in MILL’S ON LIBERTY: A CRITICAL GUIDE 62 (C.L. Ten ed., 2009).

[34] Feinberg, *supra* note 22, at 191.

[35] *See* Nicolás Quaid Galván, *Adopting the Cumulative Harm Framework to Address Second-Generation Discrimination*, 11 COLUM. J. RACE & L. 147 (2021); Katharine Gelber & Luke McNamara, *Anti-Vilification Laws, and Public Racism in Australia: Mapping the Gaps Between the Harms Occasioned and the Remedies Provided*, 39 UNSW L.J. 488 (2016).

[36] Christina Friedlaender, *On Microaggressions: Cumulative Harm and Individual Responsibility*. 33 HYPATIA 5 (2018).

that "when mediated through social systems" can harm via their cumulative and composite effects.[37] Counselors, socio-legal and criminology scholars, clinical psychologists and trauma researchers insist that the harms of microaggressions are real and need to be taken seriously.[38] But to understand how apparently "mild" abuse targeting adults can prove harmful, contextualization is critical. It is the "background condition" of structural oppression that gives these milder acts the weight of a "harm." As we wrote this essay, it was World Cup season, and the social media abuse of Black footballers in the UK, which often included the use of monkey and banana emojis, was back in the headlines.[39] By themselves, monkey and banana emojis would seem innocuous, but their use against Black players taps into a deeply racist and dehumanizing stereotype.[40] In the context, given the history behind these stereotypes and the background condition of structural racism in the UK, the use of these emojis is *harmful*, not merely offensive.

A cumulative harm framework is instructive in the context of online harms regulation, especially because most of the abuse online is ordinary, its frequent targets often belong to historically oppressed groups, and particular features of the platform environment can and do directly support the accumulation of harms.[41] While the amended UK Online Safety Bill will no longer include provisions to tackle lawful harms such as misogyny (relying instead on user controls), research has shown how

[37] Regina Rini, *The Ethics of Microaggression* 17 (2021).

[38] *See* Gordon Hodson, *Pushing Back Against the Microaggression Pushback in Academic Psychology: Reflections on a Concept-Creep Paradox*, 16 PERSPS. ON PSYCH. SCI. 932, 945 (2021); Gelber & McNamara, *supra* note 35.

[39] *See* Shanti Das, *Twitter Fails to Delete 99% of Racist Tweets Aimed at Footballers in Run-up to World Cup*, The Guardian (Nov. 20, 2022), https://www.theguardian.com/technology/2022/nov/20/twitter-fails-to-delete-99-of-racist-tweets-aimed-at-footballers-in-run-up-to-world-cup. For coverage from previous years see, for example, Cristina Criddle, *Instagram Admits Moderation Mistake Over Racist Comments*, BBC (July 15, 2021), https://www.bbc.com/news/technology-57848106.

[40] *See* David Livingstone Smith & Ioana Panaitiu, *Apeing the Human Essence: Simianization as Dehumanization, in Simianization: Apes, Gender, Class, and Race* 77 (Wulf Hund, Charles Mills & Sylvia Sebastiani, eds. 2016).

[41] *See* Ariadna Matamoros-Fernández, *Platformed Racism: The Mediation and Circulation of an Australian Race-Based Controversy on Twitter, Facebook, and YouTube*, 20 INFO. COMMC'N SOC'Y 930 (2017); Eugenia Siapera, *Organised and Ambient Digital Racism: Multidirectional Flows in the Irish Digital Sphere*, 5 OPEN LIBR. HUMANS. 1 (2019); Rosalie Gillett, *"This Is Not A Nice Safe Space": Investigating Women's Safety Work on Tinder*, FEMINIST MEDIA STUD. (2021), https://www.tandfonline.com/doi/abs/10.1080/14680777.2021.1948884.

men's violence towards women on digital platforms frequently manifest as repeated mild acts of abuse[42] that make women feel "uneasy, uncomfortable, or unsafe."[43] In fact, harm occurrence online rarely takes place as a single incident or a one-off, ephemeral, act. Content and behavior in online spaces for the most part leave traces that are often permanent, easily searchable, replicable, and scalable[44] through platforms' own design, for example via algorithmic amplification over time[45]. Safiya Umoja Noble addresses this latter point through her examination of how Google harms women of color when it returns search results for the keyword "Black girls" that portray them in overtly sexualized ways. As many civil society groups in the UK have argued (e.g., Glitch, Demos, Carnegie UK Trust), the focus of online safety regulation should not just be on content but on systems, and this systemic focus could be built into the "legal but harmful" concept. This focus on systems would allow regulators to address how platforms' own design, policies, and processes may be complicit in creating and perpetuating harm. This can encompass anything from algorithmic amplification of harmful content to facilitating the targeting of vulnerable communities through their advertising platforms (e.g., sports betting companies target groups of people with gambling problems).

Critics of liberal notions of harm have also pointed out that conceptualizations of harm that draw extensively from individualized criminal law frameworks tend to overlook social harms, which is also a key consideration within online harms regulation.[46] The effects of a "harm," like racist, sexist or ableist slights, are distributed across more

[42] *See* Emma A. Jane, *Misogyny Online: A Short (and Brutish) History* (2016); Jessica Drakett, Bridgette Rickett, Katy Day & Kate Milnes, *Old Hokes, New Media – Online Sexism and Constructions of Gender in Internet Memes*, 28 Feminism & Psych. 109 (2018).

[43] Gillett, *supra* note 41, at 2.

[44] *See danah boyd, Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications, in A NETWORKED SELF: IDENTITY, COMMUNITY, AND CULTURE ON SOCIAL NETWORK SITES 39 (Zizi Papacharissi ed., 2010).*

[45] SAFIYA UMOJA NOBLE. ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018).

[46] *See* Paddy Hillyard & Steve Tombs, *From 'Crime' to Social Harm?*, 48 CRIME, L., & SOC. CHANGE 9 (2007); Simon Pemberton, *Social Harm Future(s): Exploring the Potential of the Social Harm Approach*, 48 CRIME, L., & SOC. CHANGE 27 (2007); SIMON PEMBERTON, HARMFUL SOCIETIES (2015).

than a single individual target.[47] The concept of cumulative harm is thus inextricable from that of societal harm.[48] For example, "low-level" racism in the form of a microaggression, a "joke,"[49] or a racist stereotype constitutes a form of cumulative harm due to the existence of racial oppression (a societal harm) as a "background condition."[50] The unmitigated spread of cumulative harm sustains racial oppression as a societal harm.[51] Societal harms may be cumulative in nature, occur over the longer term and have much more complex chains (or webs) of causation.[52] The upshot of this from an online harms perspective is that when online safety regulation predominantly focuses on individual-level harm, there is a danger that the broader societal consequences of what occurs online may end up flying under the radar.[53]

A societal harm lens would allow online safety regulation to tackle racial and gender-based abuse as a problem whose impacts extend far beyond the direct targets of this abuse. At the end of 2021, when it emerged that the UK Online Safety Bill would be conceptualizing harm in exclusively individualized terms, the Joint Committee on the Draft Online Safety Bill urged for more attention on societal harms, citing testimony from experts and civil society that the harms of racial- and gender-based abuse were not restricted to individual targets and gave rise to far broader consequences—undermining principles of equality, and frequently driving members of minority groups out of political life altogether. Unfortunately, this individualistic approach appears alive and well in planned amendments to the Bill. The UK government wants platforms to provide more tools to adult users to "help them avoid" seeing

[47] *See* Nathalie A. Smuha, *Beyond the Individual: Governing AI's Societal Harm*, 10 INTERNET POL'Y REV. (Sept. 30, 2021), https://policyreview.info/articles/analysis/beyond-individual-governing-ais-societal-harm.

[48] *See* Friedlaender, *supra* note 36.

[49] Ariadna Matamoros-Fernández et al., *Humour as An Online Safety Issue: Exploring Solutions to Help Platforms Better Address This Form of Expression,* 12 INTERNET POL'Y REV. (Jan. 25, 2023), https://policyreview.info/articles/analysis/humour-as-online-safety-issue-exploring-solutions-social-media-platforms.

[50] Friedlaender, *supra* note 36.

[51] *See id.*

[52] *See* Smuha, *supra* note 47.

[53] *See* Lee Edwards, *Can the Online Safety Bill Be More Than A Toothless Tiger (Or A Facebook Flop)?*, MEDIA@LSE BLOG (Oct. 11, 2021), https://blogs.lse.ac.uk/medialse/2021/10/11/can-the-online-safety-bill-be-more-than-a-toothless-tiger-or-a-facebook-flop.

non-criminal content, but user-level tools are inadequate to mitigate societal harms.[54] Notably, a societal harm lens would also allow regulators to tackle other social ills that are not necessarily linked to systemic oppression, for example disinformation and its relation to vaccine hesitancy. The harms of misinformation information may be individual, but misinformation or commercially driven information flows, especially in high volumes, have a much broader, more distributed, and longer-term effect on a society and can undermine public health efforts with far-reaching consequences.

Social media platforms do not simply "cause" societal harm by themselves, though; they are embedded within complex media ecosystems, and they both shape and are shaped by complicated and shifting social and political environments. Because it confronts this complexity, a societal harm focus could help policymakers identify interdependencies with other policy areas and honestly acknowledge both the promise and the limits of platform-specific interventions to "solve" deep-rooted social issues.[55]

3. What Counts as Evidence of Online Harm?

Conceptualizing online harm more expansively in online safety regulation inevitably requires a conversation about what sorts of evidentiary thresholds are required to "prove" online harm. Should "background conditions" like historical and contemporary social injustices be factored into decisions around harm thresholds by platforms' online safety efforts? The Meta Oversight Board's judgment on depictions of Blackface[56] neatly captures some of these issues. The Board evaluated and eventually ruled to uphold a 2020 decision by Facebook to remove a video of Zwarte Piet containing Blackface which was shared by a Dutch user in the Netherlands. In reaching its ruling, the majority of the Board noted that Blackface "caricatures . . . are inextricably linked to negative and racist stereotypes and are considered by parts of Dutch society to sustain systemic racism in the Netherlands."[57] They also noted

[54] Silvia Milano, Mariarosaria Taddeo & Luciano Floridi, *Recommender Systems and Their Ethical Challenges*, 35 AI & SOC'Y 957 (2020).

[55] *See* Sally Broughton Micova, *What Is the Harm in Size?*, CTR. ON REGUL. EUR. (Oct. 19, 2021), https://cerre.eu/publications/what-is-the-harm-in-size.

[56] *Depiction of Zwarte Piet*, 2021-002-FB-UA, META OVERSIGHT BD. (2021), https://oversightboard.com/decision/FB-S6NRTDAJ.

[57] *Id.*

that there were "documented cases of Black people experiencing racial discrimination, and violence in the Netherlands linked to Zwarte Piet."[58] A minority of the Board, however, argued that there was "insufficient evidence to directly link this piece of content to the harm supposedly being reduced by removing it," that "while blackface is offensive, depictions on Facebook will not always cause harm to others" and that "restricting expression based on cumulative harm can be hard to distinguish from attempts to protect people from subjective feelings of offence."[59] This disagreement between Board members shows the implications of decisions to include or exclude "background conditions" as evidence of (online) harm: without evaluating the video within the context of the harmful history of Blackface and systemic racism and violence in the Netherlands, the video can be dismissed as "merely" offensive, rather than harmful (and hence fall outside of definitions of "online harm").

Individualistic evidentiary frameworks need to be adapted to capture the potential societal harms of online activity. Societal harm is more than the sum of individual harms and requires different notions of "evidence." As Nathalie Smuha has argued, there is much to be learned from environmental law in this respect—where regulators developed "societal mechanisms" to account for and preempt the accumulative and distributed nature of environmental harms like pollution.[60] These include "public oversight mechanisms to increase accountability," "mandatory impact assessments" where "impact" is understood to also include "societal impact," and giving citizens the right to request (and be granted) information (e.g., environmental data and information about government environmental policy) without having to justify the need in terms of (the risk of) individual harm.[61] In this regard, "systems and processes" regulatory proposals, like the EU Digital Services Act (DSA), hold promise insofar as they are designed to push platforms to assess how their systems contribute to "societal risks" through risk assessments, codes of conduct and crisis protocols.[62] The UK Online Safety Bill also

[58] *Id.*

[59] *Id.*

[60] Smuha, *supra* note 47, at 24.

[61] Smuha, *supra* note 47, at 1.

[62] An example of how platforms' own systems and policies can pose harm is Meta's cross-check program, which affords "additional layers of human review" to content posted by certain accounts. As the Oversight Board's policy advisory opinion argued: "The Board understands that Meta is a business, but by providing extra protection to certain users

contemplates risk assessments, but only to push platforms to evaluate how their systems pose "dangers" to children.

Conclusion: Embracing the "legal but harmful" concept with care

The term "online harm" has become pervasive in contemporary discussions about online safety regulation: but it is too frequently mentioned as though it were a settled concept, eliding the degree and nature of contestation over the term's meaning.[63] Online safety regulation proposing expansive conceptualizations of "harm" that go beyond liberal legal framings (i.e., a "legal but harmful" category) has made various players uncomfortable due to fears of over-regulation.

In response to these concerns, we have put forward two main arguments. First, resorting to narrow legal conceptions of harm to deal with all "online harms" will result in problematic blind spots. This is due to the limitations of existing legal frameworks when it comes to addressing cumulative harms to historically marginalized groups and broader societal harms as well the specificity of digital platforms and the new harms perpetrated through their networks.

Second, being expansive in definitions of online harm in online safety regulation can also pave the way for more expansive remedies for those harms. Platforms are *already* addressing "legal but harmful" content/behavior in their policies, but they often do this in a selective and reactive way in response to highly visible "public shocks"[64] and with little transparency and public oversight.[65] Regulators have shown willingness

selected largely according to business interests, cross-check allows content that would otherwise be removed quickly to remain up for a longer period, potentially causing harm." Meta's Cross-Check Program, PAO-2021-02, META OVERSIGHT BD. (2021), https://www.oversightboard.com/decision/PAO-NR730OFI.

[63] *See* work by, amongst others, linguist Sally McConnell-Ginet, who argues that the act of defining terms is "seldom just semantics" and that contests over words' meanings are worth paying attention to. *See* Sally Mc-Connell-Ginet, *Why Defining Is Seldom 'Just Semantics': Marriage and Marriage*, in Drawing the Boundaries of Meaning: Neo-Gricean Studies in Pragmatics and Semantics in Honor of Laurence R. Horn 217 (Betty J. Birner & Gregory Ward eds., 2006).

[64] Mike Ananny & Tarleton Gillespie, *Public Platforms: Beyond the Cycles of Shocks and Exceptions*. OXFORD INTERNET INST. 2 (2016), http://blogs.oii.ox.ac.uk/ipp-conference/sites/ipp/files/documents/anannyGillespie-publicPlatforms-oii-submittedSept8.pdf.

[65] *See* Nicolas Suzor & Rosalie Gillett, *Self-Regulation and Discretion*, in DIGITAL PLATFORM REGULATION 259 (Terry Flew & Fiona R. Martin eds., 2022).

to let platforms consider a range of remedies to address lawful harms that extend beyond the blunt tools of content removal and user bans, which are themselves inspired by existing criminal justice systems.[66] These include design-based approaches that introduce more user control over their online settings where appropriate, more inbuilt friction to avoid the virality of potentially problematic material, content warnings and information shelves, and reduced amplification or downranking, amongst other features. These approaches will need to be commensurate with the types of harms they seek to remedy. For example, as we argued earlier, if "low-level" racist content, such as banana emojis directed at Black people, is acknowledged as harmful, then dealing with this solely through user-level controls might help to shield individuals from direct abuse, but it will not address the societal ramifications of that abuse circulating elsewhere and hence contributing to racism as a societal harm. Having state regulation that encourages "procedural accountability"[67] can push platforms to answer for how they deal with lawful online harms and offers a "third way" between regulators ignoring lawful harms that do not meet legal thresholds and regulators resorting to unnecessary criminalization.

Regulation requiring an appropriate degree and type of platform transparency is critical to robustly test the efficacy of different measures. Indeed, the "legal but harmful" provisions in the UK's Online Safety Bill which inspired so much criticism did not prescribe remedies for "legal but harmful" content/conduct categories, but instead called for platforms to conduct risk assessments and clearly state how they would treat such risks in their Terms of Service. As a result, the Bill gave platforms the kind of latitude which can, in theory, limit regulatory overreach and over-removal (the Bill was drafted in a way that put platforms under no obligation to

[66] *See* Schoenebeck, Haimson & Nakamura, *supra* note 10.

[67] This term is used by Mark Bunting to refer to a form of accountability whereby "regulators . . . investigate intermediaries' governance procedures and incentivize them to adhere to principles of good governance, rather than to regulate their substantive rules and decisions." Mark Bunting, *From Editorial Obligation to Procedural Accountability: Policy Approaches to Online Content in the Era of Information Intermediaries*, 3 J. CYBER POL'Y 165 (2018).

Procedural accountability may involve requiring platforms to produce risk assessments, setting out how they evaluate and mitigate the risk of different lawful harms occurring on their platforms, and submitting to independent audits. Importantly, the regulator would not prescribe what platforms should do about specific cases. This provides an element of flexibility that is key to platform governance and a kind of transparency that will not solve debate over the substance of decisions, but it can at least provide the tools for a more informed and productive debate about those thorny substantive questions.

remove or even limit "legal but harmful" content/behavior, but it required them to document how they were dealing with lawful harms).

The debates around the UK's Online Safety Bill serve as an instructive example of how the "harm" concept is a site of contestation in online safety discourse. When decrying the nebulousness of the term "online harm," it is worth remembering that even outside of the online context, the term "harm" is actually "a relatively under-theorized concept."[68] The "online harms" debate is messy and is likely to remain so, but retreating to narrow legal definitions of harm is at best a partial, and we believe limiting, solution. Online safety regulation that recognizes lawful harms, and requires platforms to adopt a more structured, inclusive, and transparent approach to addressing them, offers a much-needed opportunity to engage with multidisciplinary harm frameworks and center the voices of historically marginalized groups in platform governance.

[68] Conaghan, *supra* note 27, at 321.