Attributing Responsibility to Big Tech for Mass Atrocity: Social Media and Transitional Justice

Juan Espíndola

Big Tech companies such as Meta, the owner of Facebook, are increasingly accused of enabling human rights violations. The proliferation of toxic speech in their digital platforms has been in the background of recent episodes of mass atrocity, the most salient of which recently transpired in Myanmar and Ethiopia. The involvement of Big Tech companies in mass atrocity raises multiple normative and conceptual challenges. One is to properly conceptualize Meta's responsibility for the circulation of toxic speech. On one view, endorsed by the corporation itself, Meta can be absolved from any significant share of responsibility for these atrocities because toxic speech is the speech of some (rogue) users, hosted but neither created nor endorsed by the company; if anything, Meta is responsible for failing to anticipate and swiftly remove that speech. I will argue that this view is misleading, as it misses the underlying forces crafting toxic speech. Meta's business model relies on what one might call the algorithmic capture of attention, which it achieves by manipulating its users and by creating an environment in which manipulative practices of some users thrive over others. This fact alone turns the company into a co-creator of toxic speech rather than a mere conduit of the toxic speech of others. As a result, it is safe to claim that Meta bears significant causal responsibility and sufficient moral responsibility for the dissemination of toxic speech, such that it justifies its inclusion in transitional justice processes and grounds its moral obligation to act in ways that advance these processes.

Increasingly, Big Tech companies such as Meta, the owner of Facebook, stand accused of enabling repression and human rights violations. The proliferation of toxic speech in their digital platforms has been in the background of recent episodes of mass atrocity. In a recent case, government forces in Ethiopia targeted several ethnic minorities, with Facebook propelling discourse that called for violence to be directed against them. In perhaps the most salient case, the military and nationalist groups in Myanmar used Facebook to dehumanize the Rohingya minority and to instigate genocidal acts against it.

The involvement of Big Tech in mass atrocity raises multiple normative and conceptual challenges, particularly about how to properly conceptualize social media's responsibility for the circulation of toxic speech. On one view, endorsed by Meta—which will be the focus of the paper—corporations can be absolved from any significant share of responsibility for these atrocities because toxic speech is, by and large, the speech of some (rogue) users, hosted but neither endorsed nor much less created by the company. If anything, Meta is responsible for failing to anticipate and swiftly remove such speech. I argue that this view is misleading, missing the underlying forces that craft toxic speech. Meta's business model relies on what one might call the algorithmic capture of attention, which it achieves by manipulating its users and by creating an environment in which manipulative practices of some users over others thrive. This fact alone turns the company into a co-creator of toxic speech rather than a mere conduit of the toxic speech of others. As a result, it is safe to claim that Meta bears significant causal responsibility and sufficient moral responsibility for the dissemination of toxic

Juan Espíndola (juanespindola@comunidad.unam.mx), Associate Professor at the Institute for Philosophical Research, National Autonomous University of Mexico, Mexico. He received his PhD from the University of Michigan. His research focuses on transitional justice and artificial intelligence. He is the author of Transitional Justice after German Reunification: Exposing Unofficial Collaborators *(Cambridge University Press, 2015) and co-editor of* Understanding Collaboration in Authoritarian and Armed Conflict Settings *(Oxford University Press, 2022). His most recent work has appeared in* Ethics and International Affairs, Theoretical Criminology, Journal of Social Philosophy, Critical Review of International Social and Political Philosophy, *and others.*

doi:10.1017/S1537592724001282
© The Author(s), 2024. Published by Cambridge University Press on behalf of American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

speech, such that it justifies its inclusion in transitional justice processes and grounds its moral obligation to act in in specific ways to advance such processes.

These considerations are relevant for transitional justice scholarship, a rapidly evolving disciplinary field that emerged in the 1980s and which seeks to clarify the proper ways to come to terms with the legacies of extraordinary violations of human rights. First, although the trend is reversing (e.g., Payne, Pereira, and Bernal-Bermúdez 2020), the role of corporations in social and political conflict has been neglected in extant scholarship (Jakobsen 2023; Miller 2008; van der Merwe and Brinton Lykes 2022), with its focus usually falling, understandably in part, on state actors and the main and direct perpetrators of human rights violations. When the role of corporations in mass atrocity is acknowledged and confronted, practitioners and scholars have tended to focus on cases where businesses' complicity is blatant and the perpetration of wrongdoing unambiguous. IG Farben's provision of Zyklon B for the gas chambers or Krupp's use of slave labor, both during the Nazi regime, are paradigmatic cases. More recent cases come to mind. South African banks allowed the state-owned arms manufacturer to evade the international embargo against the Apartheid regime (van Vuuren and Marchant 2022; Payne, Pereira, and Bernal-Bermúdez 2020) and several corporations in various Latin American nations shared intelligence with, or provided material assistance to, military or paramilitary groups that perpetrated acts of violence. Finally, the broadcaster *Radio Télévision Libre des Mille Collines* (RTLM) in Rwanda instigated genocidal violence against Tutsis in the 1990s. In all these cases, we can point towards clear acts of direct and intentional corporate involvement in wrongdoing. This contrasts with Meta's involvement in wrongdoing, which is of a different kind.

Before setting out, two clarifications are in order. Why focus on Facebook and not on other platforms? After all, many of them have been closely linked to toxic speech, such as YouTube, Twitter, Reddit, or TikTok, as well as others with a lower profile and geographically circumscribed, such as Parler and Gab. While toxic speech is indeed rife in these platforms, only Facebook and, to a somewhat lesser extent, YouTube have been accused by relevant stakeholders (victims, human rights associations, etc.) of being implicated in some of the crimes that are of special interest for transitional justice scholarship, such as genocide or ethnic cleansing. Similarly, both platforms, particularly Facebook, have secured a foothold in societies immersed in entrenched and sustained social conflict. Therefore, and for space constraints, it makes sense to focus solely on Facebook to discuss the harms and aspirations that unfold in transitional contexts.

The next clarifying point relates to the notion of toxic speech. By toxic speech, I mean, following Lynne Tirrell (2017, 146), speech that targets a social group with the aim and power to cause pain, inflict suffering, inspire shame and humiliation, erode one's sense of worth, damage social bonds, and ultimately undermine hope. Toxic discourse is objectionable when it explicitly incites violence, but this does not mean that those of its manifestations that fail to reach the threshold of direct incitement remain outside the focus of normative concern. In this sense, toxic speech may not explicitly and directly trigger violence, yet may be harmful by redefining the boundaries of what is considered "normal" and ultimately authorizing acts previously considered inappropriate (Tirrell 2017, 147).

Toxic speech tends to be reduced to hate speech, but this is a mistake. As I understand toxic speech, it promotes fear as much as it expresses hatred. Toxic speakers can assert, for example, that a group is planning to attack one's own group without expressing hatred, but the assertion might convince members of the latter group to condone or commit violence against the former, casting it as a self-defensive measure. Violence would then seem justified (Dangerous Speech Project 2021). Along these lines, one could claim, following Jonathan Leader Maynard (2022), that toxic speech solidifies the ideologies that promote mass atrocity. Ideology, as he understands it, is the set of guiding beliefs that underpin a kind of radicalized security politics. Following these beliefs, certain groups (of civilians) are characterized as threats, blamed for committing serious crimes (usually imaginary) and are thus made legitimate targets of "defensive" violence, and are denied common bonds of identity with the primary political community. Furthermore, these beliefs portray violence against the "threatening" group as adequate and courageous and provide an assessment according to which such violence will produce significant strategic benefits in the future, foreclosing alternative routes of action other than violence.

The paper is structured as follows. I first offer a brief overview of Ethiopia's recent civil war, which will serve as an illustration of some of the points raised in the paper. Dehumanizing and fear-inducing content disseminated via Facebook contributed to the commission of atrocities in the conflict. The case is telling for this reason alone, but it is also instructive because it shows how the relevant stakeholders and observers misunderstood the corporation's role in, and its responsibility for, the atrocities. I next develop two rival accounts of Meta's involvement in mass atrocity and its responsibility for it. One account portrays it as a neutral actor, unwillingly sucked into political and social conflicts and therefore bearing little to no responsibility for them. The alternative account reveals the real depth of its implications in massive conflict. I then lay out what this discussion entails for transitional justice processes, focusing on three

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

aspirations common to these processes: truth-seeking, guarantees of non-repetition, and criminal trials.

Toxic Speech and Mass Atrocity: Ethiopia as an Example

Social media platforms have made themselves complicit in human rights violations in some obvious and uncontroversial ways. To mention a recent example, consider the case of VKontakte (VK), Russia's local substitute for Facebook, which is blocked in the country. After VK emerged as an independent tech start-up, its popularity in Russia surpassed YouTube and near paralleled Telegram. Progressively, however, it came under the control of the Kremlin, as its original owners were forced to sell their shares to a pro-Kremlin oligarch. It now takes orders from authorities concerning what content makes it onto the platform and what does not when human rights violations involving the Russian invasion of Ukraine are at stake (Meaker 2022). This kind of deliberate assistance to a wrongful party—an unjust aggressor—turns VK into a willing co-conspirator.

But social media's complicity in wrongdoing is elusive when the platform's contribution involves tolerating toxic speech rather than providing explicit assistance to perpetrators. The case of Ethiopia's recent civil war vividly illustrates Facebook's contribution to mass atrocity via the proliferation of toxic speech on its platform and raises important questions about the company's responsibility. The situation in Ethiopia differed from that of Myanmar, mentioned earlier. In the latter case, Meta might have argued that the harmful effects of toxic speech were unprecedented and, hence, difficult to anticipate. However, in Ethiopia, given Myanmar's precedent, the corporation could not claim this kind of non-culpable ignorance, assuming it was ever credible. This circumstance renders Ethiopia an appropriate case for considering platform responsibility.

The most recent of Ethiopia's long list of internecine struggles erupted in 2020, pitting the Ethiopian government against Tigrayans and other ethnic groups. The conflict gave rise to well-documented instances of crimes against humanity, ethnic cleansing, weaponized rape, and genocide, among other crimes perpetrated primarily by Ethiopian government forces. The civil war concluded in 2022 with a peace agreement and the prospect of a transitional justice process, one with no indication of confronting Meta's role in the civil war.

Facebook, with 6 million users in Ethiopia, was accused of providing a platform in which narratives dehumanizing and stigmatizing Tigrayans spread widely. For instance, Tigrayans were commonly accused of treason. In July 2021, in a Facebook post that remained on the platform long after the conflict had initiated, Prime Minister Ahmed referred to Tigrayan rebels as a "weed" that must be pulled out. Similarly, in a widely engaged and shared post, a popular public figure, appearing frequently in state television, echoed this call to violence exhorting his more than 120,000 followers across the country to assassinate members of the Tigrayan ethnic group. "The war is with those you grew up with, your neighbor …. If you can rid your forest of these thorns … victory will be yours" (Zelalem and Guest, 2021). Another ethnic minority, the Qemant, was also targeted by government forces and allied militias. In September of 2021, a Facebook post that received hundreds of reactions falsely alleged that terrorists from a Qemant village hijacked a bus, resulting in the assassination of two people. The village was pillaged and razed for several days (Zelalem and Guest, 2021). Finally, relatives of murdered Tigrayans accused "online activists" like Solomon Bogale of fanning violence. With more than 86,000 Facebook followers, Bogale posted images of himself carrying an assault rifle, accompanied by statements praising a vigilante group of nationalists from the rival Amharan ethnic group, or calling to violence along these lines: "We need to cleanse the region of the junta lineage [Tigrayans] present prior to the war!!" (Jackson, Kassa, and Townsend 2022).

The bulk of (self-)criticism for Facebook's involvement in the egregious human rights violations revolves around its failure to purge content that was demeaning of, or that incited violence against, Tigrayans and other groups. Its meager moderation tools were singled out as the source of the problem: human content moderators lacked knowledge of all local languages relevant to the conflict (Amharic, Oromo, Somali, Tigrinya); content moderation had been outsourced to a company in Kenya with poor labor standards and permanently under pressure from Meta to put speed ahead of thoroughness in its review processes; and users had failed to alert the platform about the proliferation of toxic speech, perhaps owing to their lack of digital literacy (in the sense that reporting interfaces might have been confusing to them because they were still unfamiliar with the platform). Many observers also noted that the main problem had been one of inadequate *automated* content moderation. In fact, according to the so-called Facebook Papers, leaked by whistle-blower Frances Haugen, before violence erupted in Kenya, employees at Meta were aware that the data collected from users to understand problematic content ("signals") were proving inadequate to monitor the situation. In response, and notwithstanding its experimental status, the company put a novel content moderation technique to the test, which identified patterns of behavior consistent with malicious activity rather than using specific words or phrases to directly identify hate speech or misinformation. The efforts were to no avail (Zelalem and Guest 2021; Elliot and Cameron 2022; Jackson, Kassa, and Townsend 2022).

This diagnosis set the parameters for Meta's proposal for corrective action. In subsequent public statements,

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

Meta reaffirmed its commitment to improving content moderation tools, ramping up its efforts by *inter alia* making sure moderators were proficient in all local languages, and vowing to improve automated moderation tools. Meta claimed to have strengthened its capability for removing content promoting hate speech, disinformation, and incitement to violence. For instance, while the conflict was still unfolding, Meta claimed to have removed "content that provides material support towards [violent] groups or praises the violence they commit;" crafted "a more extensive list of slurs across the four main Ethiopian languages;" and purged "content that contains claims about individuals being spies, traitors, or informants." (Meta 2021a). In this same spirit, when a post falsely accused the Tigray People's Liberation Front and ethnic Tigrayan civilians of committing atrocities in Ethiopia's Amhara region and inflamed ethnic tension by asserting that "we will ensure our freedom through our struggle," Meta's Oversight Board (its highest decision-making body in some respects)[1] instructed the company to remove it. Initially, the company had reached conflicting resolutions about what to do with the post, but the Oversight Board determined it violated Violence and Incitement Community Standards—but not the Hate Speech Community Standard (Oversight Board 2021; Oversight Board 2022).

Such criticisms and actions are partly on point, but they miss considerations that are critical to elucidate the full extent of Meta's causal and moral responsibility. They are inadequate because they misrepresent the nature of the technology that social media is, taking it to be a platform that merely "hosts" the speech of users. As a result, they wrongly characterize the corporation's responsibility solely in terms of its failure to promptly detect speech that promotes hate or incites violence through moderation tools such as content takedowns, degradations, or account suspensions. As I now will argue at length, this assessment does not reach deep enough into the source of the problem and is, therefore, unreliable as the basis for responsibility attribution.

The Algorithmic Capture of Attention: An Objectionable Business Model

What kind of contribution do social media platforms make to the proliferation of toxic speech? On its face, platform companies make an involuntary contribution to it. Like the manufacturers of megaphones whose buyers use them to insult passers-by and perhaps foment public riots, they are the vehicle through which some of their users propagate toxic speech, even if, in most cases, platform companies do not endorse the content of toxic speech. Since these companies cannot always fully anticipate how other agents will make use of their services, it could be argued that their *causal* responsibility for toxic speech is limited and their *moral* responsibility diminished. Even if platform companies could anticipate the appearance and dissemination of toxic speech, their responsibility would lie in their recklessness in providing access to the platform to its creators. Call this the neutrality view of toxic speech; I argue against it. The central claim is that the contribution of social media platforms to disseminate toxic discourse is much more significant, both causally and morally, than the neutrality view admits. Platforms are not a "neutral" technological vehicle of the toxic discourse created by users, and their responsibility for the proliferation of that discourse is by no means secondary or derivative.[2] On the contrary—and this is the critical point—the contribution of digital platforms to toxic speech must be traced back to a business model—call it the algorithmic capture of attention—that, through automated *manipulation*, organizes information and creates user habits, which in turn are likely to harbor toxic speech. By manipulation—online or offline—I simply mean a deliberately surreptitious influence exerted to control people (Noggle 2022). It is no exaggeration to claim that this contribution turns platforms into co-authors of toxic speech. Call this the manipulative view of toxic speech.

Before making the case *against* the neutrality view and *for* the manipulative view of platform responsibility, I shall clarify what I mean by platform responsibility for toxic speech. Like responsibility more generally, it has a causal and moral dimension. Causally, it comes in degrees and refers to the various kinds of contributions that platforms make to the formation of toxic speech. These contributions range from the failure to detect or contain toxic speech, to more active interventions such as promoting or creating it. Both action and inaction on the part of the platform can, of course, be causally relevant for the proliferation of toxic speech: which kind of intervention (or lack thereof) is more causally efficacious is a matter to be determined contextually. Morally, platform responsibility is also scalar and refers to the amount of blame that can be attributed for these wide-ranging contributions. At one extreme, blame can be assigned for omissions that resulted in failing to remove toxic speech, which in turn might stem from recklessness or culpable or non-culpable ignorance of its effects. Greater blame can be assigned for deliberately promoting toxic speech, particularly when the intent is to harm a person or a group (e.g., for ideological reasons) but also when there is no such intent and harm is a side effect arising from the pursuit of a different goal, like profit-making. Greater or lesser blame for toxic speech can also be a function of its ensuing harm. Failing to detect and remove a particular kind of toxic speech may be more blameworthy than actively promoting a different kind when, say, the former causes much greater harm than the latter.

With this clarification in mind, I now return to the neutrality view. To unpack this view, we can draw a parallel between the paradigmatic case of a manufacturer selling arms to groups that will utilize them to wrong

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

others and the case of platform companies providing access to groups that will use the platform to voice hatred and incite violence (Howard 2022). An arms seller is not complicit with his buyers solely if he endorses their intention to inflict harm. The seller would also be complicit, although to a lesser degree, of course, if he was unaware of the buyers' nefarious plans, *but* could (or should) have anticipated them (Lepora and Goodin 2013) because while admittedly the seller did not intend the harm, the sale betrays a disregard for the risks of harm being inflicted on others.[3] Similarly, for Meta to be complicit, the company need not support the goals of users who use the platform to disseminate toxic speech, as VK does when it shares misinformation or distorts accurate facts at the behest of authorities in the Kremlin. Suffice it that it fails to foresee the risks entailed in how these users will (mis)use the platform and as a result grants them access to it, which wrongs third parties.

The problem with this widely shared account of platform responsibility is that it locates the *source* of toxic speech outside of the platform and represents the latter as a mere transmission band. This is the sense of the ubiquitous expression of "user-generated content." Content in the platforms, it is alleged, is not produced by the (neutral) platforms but by their wide array of users (ordinary users, local political actors, transnational and foreign actors) who circumvent the platform's restrictions. Sometimes it is "bad actors" who intend to promote toxic speech; they may want to create political instability or sow generalized distrust. This "coordinated inauthentic behavior," as the tech jargon describes it, is the main engine of toxic speech. Put briefly, toxic speech is neither the making of platforms nor their fault. If anything, platforms are responsible for non-culpably omitting to remove the content that bad actors produce.

The reasoning used by a U.S. Court of Appeals to dismiss the 2019 judicial case *Force v. Facebook, Inc.* is a good example of the neutrality view.[4] In this case, the plaintiffs argued that the corporation had aided and abetted the organization Hamas because it had failed to remove content that celebrated and encouraged violence against Israeli citizens, and because its algorithms "directed such content to the personalized newsfeed of the individuals who harmed the plaintiffs." Put differently, Facebook[5] had not been an innocent vessel of harmful content: the corporation had also "developed" the content and should not, therefore, gain the immunity extended by Section 230(c)(1) of the Communications Decency Act. This piece of legislation is the backbone of so-called intermediary liability. The Court of Appeals was unpersuaded. It insisted on giving immunity to Facebook, positing *inter alia* that Facebook did not "develop" the content in question. An organization develops content, so claimed the Court relying on a "material contribution test," when it gives specific instructions on how to craft corrupt publicity that encourages users to act in a certain way (e.g., to buy fraudulent products). The test, in other words, distinguishes between merely displaying illegal content and bearing responsibility for what makes the displayed content illegal. Facebook, which the Court portrays as a "neutral intermediary," did the former, not the latter. The majority opinion concluded: "Merely arranging and displaying others' content to users of Facebook through [its] algorithms …. is not enough to hold Facebook responsible as the 'develop[er]' or 'creat[or]' of that content."

The manipulative view is less sympathetic to Meta's alleged neutrality vis-à-vis the content that circulates on the platform precisely because it regards Meta as a co-creator of that content. It contends that "user-generated content" is a misnomer that obscures the platforms' role in content generation, which I will now explain. At any given moment, Facebook users could be served unfathomable amounts of information of all sorts—inane and dangerous, toxic and non-toxic. This information needs to be organized and distributed somehow, lest the sheer quantity overwhelms users and renders the platform useless. This is what Josh Simons (2023, 104-133) calls the problem of abundance. To tackle it, Facebook undertakes to organize and distribute content based on a set of criteria of *its choosing*. Simons calls our attention to four "points of choice": the corporation creates a ranking system with an underlying set of values that prioritizes the social—the tribal—over the public, and engagement over quality; it chooses to optimize for "meaningful social interaction" (that is, active forms of engagement rather than more passive ones); it establishes what, if anything, counts as toxic; and provides guidelines for people who label the examples to train the algorithmic models. In effect, all these choices give the corporation the ability to promote or demote particular items of information.

It is precisely Meta's power to make these choices that turns it into a co-creator of toxic speech. These choices are a kind of editorial judgment or curation. As we have seen, they are not dictated by first-order considerations such as an ideology, but by second-order considerations related to profit: whatever the ideological underpinning of the content, it will be promoted if it can be monetized. To deny that this kind of curation amounts to editorializing would be analogous to denying authorship to the artist who creates a collage because the hundreds of photographs she collected from magazines and collated to the work were taken by someone else.

Now, it would be disingenuous to claim that Meta organizes and distributes content solely to solve the abundance problem. By intervening in the exchange of information, it becomes what Seth Lazar (2023a, 2023b) rightly calls an algorithmic intermediary, wielding unprecedented economic and political power (Coeckelbergh 2022, 18-19, 97-98), which it has used to create a novel

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

business model.[6] The underlying goal of the model is to make users interact with the platforms for as long as possible. The most obvious mechanism to achieve this is to make an addiction out of its use, a strategy with deep ethical consequences that include the exploitation of the users themselves (Bhargava and Velasquez 2021). Another strategy to seize users' prolonged attention, and the one I wish to focus on here, is to encourage the most effective type of speech to "hook" people. Toxic speech has this quality, as it tends to elicit strong emotional responses (anger, indignation, resentment, and so on) and the platform cultivates it by manipulating users through the architecture of its pages, with so-called affordances (page design properties that invite the user to interact with technology in a certain way)[7] and with algorithmic personalized recommendations.

Surely the most concerning tool of manipulation is algorithmic recommendation systems, which serve content based on a prediction of users' preferences. Recommendation systems have rightly been identified as playing a central role in the formation of digital pathologies such as polarization, especially when they create "epistemic bubbles" or "echo chambers" (Nguyen 2020). Research in the social sciences yields mixed results relating to the question of whether and to what extent algorithmic recommendation produces polarization and civil unrest.[8] Beyond polarization, and more relevant to this paper, algorithmic recommendations encourage or enable other pathologies such as "algorithmic deviancy amplification" (Wood 2017) or radicalization (Cassam 2022; Alfano, Carter, and Cheong 2018). Algorithmic deviancy amplification is a process through which the user who has interacted with content that validates or promotes "antisocial" behavior or who joined online groups that do so will subsequently continue to be exposed to such content, perhaps even more extreme content. These phenomena produce "criminogenic positive feedback loops of information consumption" (Richards and Wood 2020, 115), which in turn create or reinforce criminogenic attitudes and beliefs, in a kind of "algorithmic drift" (Richards and Wood 2020, 115).[9] Radicalization, in turn, refers to the inclination to no longer regard what once were considered extremist policies as such; in the process, "what would once have been unthinkable … becomes not only thinkable but politically feasible" (Cassam 2022, 168). Radicalization, then, "normalizes extremist policies and makes it possible for people who do not think of themselves as extremists to adopt them" (Cassam 2022, 168-9).[10]

As Alfano, Carter, and Cheong (2018) note, recommendation systems convey to users the sense that these systems (Facebook's Newsfeed, Google's search engine, etc.) know what users are thinking or looking for. To some extent, recommendation systems predict what the preferences of users will be, but, importantly, since preferences are *malleable*, they also *shape* these preferences rather than simply predicting them. Note that manipulation via recommendation systems in digital platforms is significantly stronger than manipulation exerted through traditional and nontechnological media and, therefore, harder to escape. This is because, unlike the latter, the former can remove the friction generated by rival items of information by never presenting them to users, who in the current informational ecosystem struggle to find reliable sources of information. Similarly, in traditional media, manipulation is exclusively top-down: it works by providing the categories through which consumers perceive the world. Algorithmic manipulation combines top-down with bottom-up manipulation, which is based on profiling users and predicting/shaping their preferences. From one recommendation to the next, bottom-up manipulation opens a path that users take without knowing its endpoint, potentially leading to self-radicalization (Alfano et al. 2020). The interaction between top-down and bottom-up manipulation forms a feedback loop that is harder to escape than unidimensional forms of manipulation, typical of traditional media (Alfano, Carter, and Cheong 2018).

In short, to the extent that the platform-engineered preferences and habits are prone to produce toxic speech, platform companies bear significant *casual* responsibility for toxic speech. This is not to deny that "rogue" actors have an important role in disseminating toxic, as the neutrality view posits. In fact, as these actors manipulate other users, platforms provide them with the tools to refine (e.g., by offering personalized recommendations that target users' cognitive weaknesses) and amplify (in the sense of expanding their reach to broader audiences) their manipulative interactions. In the specific case of Meta, then, the company is responsible for building a tool, Facebook, designed to manipulate, *as well as* for enabling some users to engage in manipulative interaction. Both forms of manipulation compound to form toxic environments within the platform.

The next question is, are platforms not only causally, but also *morally* responsible for toxic speech? The manipulative view posits that they are to some degree. Meta is fully responsible, morally, for the first form of manipulation just alluded to (i.e., *its* own manipulation). This is not because manipulation is intrinsically wrong (it is not[11]) but because the manner and the purpose for which Meta and similar platforms manipulate makes it decisively wrongful. First, while manipulation need not compromise a person's autonomy, or compromise it only partially or temporarily to strengthen it in some other dimension or in the long run, platform manipulation is different because it commonly exploits users' heteronomy intentionally *to gain some advantage from them.* And even if, on occasion, platforms do promote user autonomy, for instance by providing information that helps them achieve their goals,

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

the deeper problem is that they instrumentalize the autonomy of users for their own ends (Engelen and Nys 2022). Relatedly, while some kinds of manipulation may advance the manipulee's interests, platform manipulation may prompt users to endorse and maintain beliefs and desires from which they feel alienated (Engelen and Nys 2022; Williams 2018). It is surely not in the best interest of users to be constantly enraged, to sustain their enragement based on dubious or unverified information, and to act upon it. Finally, while manipulation may be acceptable if it does not override the ability of a person to revise their beliefs or conduct, platform manipulation may be hard to surmount. As we have seen, its mechanism of influence is insidious and overwhelming, preventing users from reconsidering beliefs they might have formed in their interactions with the platforms. These mechanisms may exploit users' vulnerabilities and cognitive biases or undermine their ability to make choices by making other choices exceedingly alluring and tempting (Susser, Roessler, and Nissenbaum 2019). It is no exaggeration to claim, as Wildman, Rietdijk, and Archer (2022) do, that platforms govern the affective lives of users, as when they seek to trigger strong negative affective reactions toward certain individuals or groups. Platforms may also fail to adequately alert users of the epistemic risks they will run into when they navigate the platform, such as being presented with disinformation masquerading as verified news, thereby making it hard for users to revise their ideas (Gunn 2022).

Let us turn now to the second kind of manipulation, that is, the manipulation of particular users, which Meta enables. Meta is partly, but not primarily, morally responsible for it. Since Facebook allows for more refined and widespread manipulation, it is partly responsible for it. Yet had these specific manipulative acts not existed in the first place, Meta would not have amplified them and offered tools for their refinement. Hence, the corporation bears only partial responsibility for them.

I must also clarify that it does not follow from the preceding arguments that Meta bears full responsibility for mass atrocity or other outcomes that are rendered more likely by these two forms of manipulation. The corporation neither intends to instigate genocide nor downplay it. Meta must not be attributed full responsibility for these outcomes because they are contingent on the presence or absence of additional factors ("factors of escalation," such as social fractures, the machinations of political actors, etc.; and "factors of restraint," such as the promotion of tolerance by civil society institutions like churches [Straus 2012]) that are beyond the control of Meta. Meta does bear some responsibility for failing to duly acknowledge these background conditions and consequently failing to adopt a cautious approach while supplying its service. But this is not enough to blame it to a high degree for their occurrence. It might be contended, counterfactually, that atrocity would have taken place irrespective of Facebook's contribution to it. Even if this were true (the claim is unverifiable), Meta is partly responsible for failing to extricate itself from the circumstances that did produce mass atrocity.

In sum, platforms do not simply host the toxic speech of others. Even authors who are fully cognizant and critical of Facebook's manipulative strategies for capturing users' attention put the onus squarely on political actors when it comes to atrocity-inducing toxic speech. Thus, Siva Vaidhyanathan (2018, 197) writes that the platform "allows authoritarian leaders and nationalist movements to whip up sentiment and organize violence and harassment against enemies real and imagined." However, he concludes that "Facebook does not favor hatred. But hatred favors Facebook" (Vaidhyanathan 2018, 197). As I have argued in the section, this might be too hasty a conclusion. After all, Facebook's choices for organizing information make a heavy contribution to the production of hatred. At any rate, I have outlined a variety of grounds for attributing significant causal responsibility and some degree of moral responsibility to Meta. They are sufficient to justify the incorporation of companies like Meta into transitional justice processes, if only to explore whether it inadvertently inculcates violence-inducing toxic habits and whether it does enough to curtail the detrimental effects of Facebook and to extricate itself from users' wrongdoing. To this point we now turn.

Transitional Justice in the Age of Digital Platforms

Why does defending the manipulative view of Meta's responsibility for toxic speech matter for transitional justice processes? This section addresses this question. The gist of the argument I offer in response is that, given the responsibility for toxic speech borne by Meta, it has a moral obligation to act in specific ways to advance transitional justice.

Before developing this claim, however, let me explain why probing the question of the nature of Meta's responsibility for toxic speech is particularly germane to societies aspiring to transition away from mass violence. While all societies experience the harms of toxic speech (the January 6 insurrection in the United States is a case in point), these harms are exacerbated in transitional societies, given their political and social conditions. What are these conditions? For one, wrongdoing that unfolds in transitional contexts is of a distinctive kind. It is not rare or exceptional; instead, it has been normalized, becoming a basic fact of life for members of whatever group is targeted, who then need to orient their conduct around the possibility of being its victims. It is collective in the sense that individuals are targeted as members of a group by perpetrators acting in a group. It is political insofar as perpetrators of wrongdoing typically include state actors (police, security forces). Furthermore, wrongdoing occurs against a background

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

of pervasive structural inequality (institutions across the board treat individuals unequally where political and civil equality should prevail) and political uncertainty (it is unclear where ultimate authority lies and whether the political community is even viable; Murphy 2017).

Social media platforms, as conveyors of toxic speech, can make these conditions even more acute for many reasons (some of which are discussed in Murphy 2022). Through toxic speech, these platforms can normalize wrongdoing, contribute to the orchestration of ill-intended collective action and the formation of stereotypes about groups, and be instrumentalized by political actors. Furthermore, in the presence of pervasive structural inequality, access to, and immunity from the noxious effects of social media platforms will be unevenly distributed. Privileged social groups may leverage social media platforms to their advantage, and they will be unlikely targets of toxic speech, at least of a particularly intense kind. By contrast, the ability of less privileged groups to have a voice or some influence in the platforms will be greatly curtailed, their views will be marginalized or absent, and they will be more likely targets of toxic discourse. Finally, if, in transitional societies, political authorities are weak, they will do little to curb platforms, which will then be left to self-regulate in the absence of incentives to reign in toxic speech promptly. If political authorities are strong but authoritarian, they may feel tempted to instrumentalize the platforms to reach their ends, which can include the promotion of toxic propaganda.

Advancing an appropriate interpretation of platform responsibility bears relevance to the shape of transitional justice processes. The ultimate aspiration of transitional justice is social transformation. The manipulative view of platform responsibility offers better grounds than the neutrality view for materializing this aspiration ambitiously and robustly. In other words, understanding platform responsibility along the lines of the manipulative view gives transitional justice the appropriate focus, depth, and direction, as well as provides guidance for resolving the potential objections and moral conflicts that the process raises. To elaborate on this claim, I consider three interrelated transitional justice aspirations: truth-seeking, non-repetition, and criminal justice.[12]

Truth-seeking: focusing on the black box of corporations rather than the deeds of "bad" actors. The aspiration of clarifying the truth about the causes, the agents, and the consequences of social and political conflict is central to transitional justice. Truth commissions—institutions central to achieving this aspiration—tend to deal preponderantly with the structural causes of the conflict. The examination of these causes is presumed essential to shed a bright light on the causes of mass violence (Hayner 2010). Along these lines, a common assumption under which truth commissions operate is that, in the short term, identifying and confronting the actions of particular individuals is not always possible, due to material or budgetary limitations; desirable, due to pragmatic considerations; relevant, since the focus on this or that particular actor would distract from the examination of systemic/institutional causes; or simply is the task of other institutions.

The manipulative view of Meta's responsibility lays the conceptual and normative groundwork for thorough scrutiny of the computational power and digital architecture of social media platforms along with the expectations set by truth-seeking processes. The goal of these processes in the present context would be to show how some of the key features of digital platforms like Facebook promote certain content and demote others, thereby co-creating toxic discourse. Their algorithms are usually inaccessible to most people. This opacity is partly a consequence of the technical complexity of the algorithms, which are sometimes not even fully understood by their developers. But the opacity of the technology is also the result of a range of legal and political forces ranging from intellectual property law to the ability of tech companies to capture political institutions through lobbyists, construct close relations with public authorities, centralize decision-making within companies themselves, and even frame their *modus operandi* as advancing valuable goals such as freedom of expression (Zuboff 2019). It is precisely the legitimacy of such forces that must be called into question to clarify the truth. Any policy that fails to require transparency about the corporation's algorithms is insufficient under the metrics of establishing the truth. Proprietary algorithms cannot remain in the dark if truth-seeking is to be attained. As Monsees and Srivastava argue, Meta and other Big Tech companies are not representative bodies, and yet they make decisions that affect millions and sometimes billions of people, which raises the question of how they can secure legitimacy (Monsees et al. 2023, 18; see also Srivastava 2021; Benn and Lazar 2022). In the wake of atrocity, the question of who grants them that right and under what conditions can be taken up by truth commissions.

By contrast, the neutrality view of Meta's responsibility does not get us very far in the search for truth. Under this conceptualization, clarifying the truth would be reduced to elucidating the role of social media platforms in the production of toxic environments. To be scrutinized under this conceptualization, for example, is the formation within platforms of echo chambers: digital spaces deliberately created by political or social ("bad") actors, to discredit rival opinions, which reduces the plurality of information sources, including those that could counteract the toxicity of certain discourses. This scrutiny may illuminate some of the supra-individual sources of conflict but would not provide much leeway to direct the search for truth towards the corporation's "black box," that is, towards the mechanisms for the algorithmic capture of

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

attention that underpin its business model and encourage toxic discourse.

Guarantees of non-repetition: prioritizing pre-emptive policies like algorithmic redesign rather than reactive policies like content moderation. It is commonplace and relatively uncontroversial to hold that, without a far-reaching reform process that transforms the institutions that caused or did not prevent extreme conflict, there are no guarantees that cycles of violence will end or subside, and transitional justice will hardly achieve its goal of social transformation. The institutional overhaul in question must include corporate institutions.

The guarantees of non-repetition usually invoked in scholarly and public discussions tend to revolve around the ability of platforms to moderate content appropriately and to curb the ability of bad actors to wreak havoc. These guarantees fall under the remit of human rights due diligence: the processes of investigating how corporate activities will impact the basic rights of others. This is a commitment enshrined in many local and regional jurisdictions, as well as in international frameworks like the UN's *Guiding Principles on Business and Human Rights* (2011). Meta endorses these guarantees of due diligence, vowing to "pay particular attention to the rights and needs of users from groups or populations that may be at heightened risk of becoming vulnerable or marginalized" (Meta 2021b). Some of the specific steps the company takes to honor due diligence are "efforts to help prevent election interference; assess and improve our content review operations; ensure platform integrity; foster responsible innovation; implement and uphold privacy principles; assess how we respond to government data disclosure requests; uphold the highest supply chain standards; and understand our social impact" (Meta 2021b).

These measures, which reflect the neutrality view of Meta's responsibility, are important but limited. For one, relying solely on them is unwise because fake social media accounts, which spew the kind of toxic speech of interest here, are increasingly hard to detect, as a recent intelligence leak revealed (Menn 2023). For another, the demands these guarantees impose on corporations are easily avoidable. Since they entail, to some extent, gathering information about the intentions and plans of some users, and since collecting such information is challenging, Meta can argue, plausibly, although not always genuinely, that no amount of investigation would allow it to anticipate infallibly the misuse of the platform; a loophole opens here for the evasion of its obligation.

There is a larger problem, however. With a few exceptions, most of these steps take stock of how Facebook will *react* to the actions of some of its rogue users. Consider content moderation, understood here as those systems that identify and organize the content based on certain criteria, with consequences regarding its permanence and hierarchy on the platform, such as its removal, degradation, and geoblocking, among others (Gorwa, Binns, and Katzenbach 2020; Gillespie 2018). The task of moderation is characterized as unveiling users who abuse the platforms and putting a halt to them. Similarly, consider platform integrity: its goal is, in the words of a former member of Facebook Civic Integrity, to defend the platform "from attackers who have found and learned to abuse bugs or loopholes in its rules or design" and "to systematically stop the online harms that users inflict on each other" (Massachi 2021). There are several problems with these *reactive* measures. As discussed earlier, "coordinate inauthentic behavior" can be undetectable. More importantly, though, these measures, on their own, do not address the root problem, which is that the source of toxic speech is, to a significant extent, internal to the platforms.

The manipulative view of Meta's responsibility yields guarantees of non-repetition that are compatible with these tools but significantly more ambitious and difficult to evade. What these guarantees impose on platform companies is not merely that they refine and improve their ability to detect and stop the deeds of bad actors who are bending the platform toward their ends. The demand involves overhauling platform architecture and algorithmic design in such a way that the recommendations of algorithms do not emanate (exclusively) from the pursuit of greater involvement of, and interaction between, the users. In other words, the business model must delineate strategies that capture the attention of users in permissible ways or not at all.

I cannot offer an exhaustive menu of the guarantees of repetition that would be called for, especially because digital platforms evolve constantly and rapidly, and because there is controversy around which measures work effectively to reduce algorithmic influence (see for example Guess et al. 2023a; Guess et al. 2023b; Nyhan et al. 2023). I mention some relevant guarantees. One of them relates to reducing the ability of actors who deliberately escalate the conflict to reach large audiences by using algorithms like PageRank, which makes the virality of content dependent on the trustworthiness of the source (Stray, Iyer, and Larrauri 2023, 26). Platforms could design their recommendation systems to introduce some degree of randomness or "noise" into their recommendations to expose citizens to a more heterogeneous menu of choices (Reviglio 2017; Sunstein 2017), and in the best scenario foster "constructive conflict" by algorithmically favoring content with a cross-partisan appeal (Stray, Iyer, and Larrauri 2023, 26). Other measures include banning targeted recommendations altogether, which would upend the entire business model or, less drastically, giving users control over recommendation algorithms by guaranteeing their right to block the use of some or all of their data to feed algorithmic ranking systems (Keller 2021, 32). Yet other measures include subjecting algorithms to digital clinical trials by a government agency

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

(Groth, Nitzberg, and Russell 2019); creating a Platforms Agency that requires Big Tech to justify their algorithmic choices (Simons 2023, 183-211); building democratic digital environments (Forestal 2022, 177-189); or fostering competition between platforms, which would strengthen the ability of users to "choose" from different algorithmic recommendation systems, interface designs, and even content moderation rules (see Keller 2021, 34, but others dispute the idea that competition will improve content moderation). In many ways, the regulative architecture of the European Union (2022), including the Digital Services Act and Digital Markets Act, provides a model for other jurisdictions.

All these measures would encounter implementation obstacles in any society but come with additional challenges in transitional ones. It is unclear whether transitional societies can produce the kind of competitive platform market discussed earlier since they may lack the legal tools to create them (e.g., the juridical means to force companies to share their data) as well as the technological infrastructure that can sustain such competition. Similarly, arguing for user control over personal data as a means of "taming" recommendation algorithms presupposes that users have some degree of digital literacy—in the limited sense that they have a minimal understanding of the multiple manners in which they can interact with digital platforms, and which of these work to their advantage. Users in transitional society may not have such competency. Recall the case of Ethiopia, where users failed to alert Facebook content moderators of the proliferation of toxic speech, presumably because of their lack of familiarity with the platform's interface. At any rate, the point to underscore is that guarantees of non-repetition must include policies that operate prior to the appearance of bad actors and toxic content.

To conclude this subsection, let me clarify that the foregoing discussion should not be taken to suggest that the reactive approach of removing content is superfluous. It is not, and in fact, the manipulative view of platform responsibility offers stronger grounds than the neutrality view for defending content moderation against one of its most common objections. According to this objection, the benefits of eliminating toxic speech through content moderation come at the cost of unduly restricting users' freedom of expression. Much of the strength of the free speech objection to content moderation depends on viewing Facebook merely as the provider a forum where opinions are disseminated, without the forum having any influence in shaping these opinions. If, instead, we view the platform as a manipulative co-creator of toxic discourse, then the weight of the free speech objection is weakened, if not fully removed. If the expressions of users in digital platforms are partly induced by the platform itself through manipulation, then these expressions embody to a lesser degree the values such as personal autonomy that the right to freedom of expression seeks to protect (Sahebi and Formosa 2022; Simons 2023, 192).

Criminal trials: Focusing on the foreseeability of algorithms' consequences rather than on the foreseeability of users' actions. The debate about the desired scope of criminal trials in transitional societies is one of the most contentious in transitional justice scholarship and practice. The considerations at stake include the classical worry that criminal prosecutions may imperil political stability or peace (Nino 1997) or the more recent concern about the opportunity cost of investing in criminal prosecutions, always onerous, rather than in areas of greater socioeconomic urgency (Swenson and Kniess 2021). However, it is important to stress that what underlies this transitional justice aspiration is the desideratum of signaling a break with the past by publicly repudiating the actions of wrongdoers and holding them accountable.

Before directly addressing the proper grounds for attributing responsibility to corporations in the context of transitional justice processes, I provide some background context. Attributing criminal responsibility to corporations in transitional contexts raises interesting questions about the basis for liability for prosecution. In prior criminal trials involving corporate wrongdoing, limited as they have been locally and internationally, individual *intentions* were central. Two paradigmatic international trials are the case of Nazi propagandist Julius Streicher in the aftermath of World War II and that of executives and co-founders of RTLM in the wake of the Rwandan genocide. In these cases, the International Military Tribunal and the International Criminal Tribunal for Rwanda, respectively, were adamant about the importance of examining the intention of the defendants for prosecution. As editor of the newspaper *Der Stürmer*, with its motto *Die Juden sind unser Unglück* (the Jews are our misfortune), Streicher had actively disseminated modern "blood libels." Similarly, Rwandan broadcasters openly called for the extermination of Tutsi "cockroaches." Both defendants unambiguously endorsed the genocidal plans of their superiors or co-conspirators. We could understand these two emblematic cases as *incitement to genocide.*

Big Tech companies hardly ever intend for their platforms to be used to spew toxic speech. In the case of Ethiopia, discussed earlier, Meta and its members (from its CEO to its content moderators) did not endorse the values or strategies of public authorities or users who targeted and degraded ethnic minorities. Rather than seeking to ground criminal responsibility on accusations like incitement to genocide, with the ensuing intent requirement, the more important question is whether Meta could have (or should have) *foreseen* ill intent on the part of those who used its services. Such an approach makes the requirement of intent in the criminal charge unnecessary, replacing it with that of knowledge or even predictability (*foreseeability*, for short). In international criminal law,

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

the articulation of the knowledge standard can be traced back to one of the so-called industrialist cases, particularly the Zyklon B case, in which two members of the company that produced the gas used in extermination camps were found guilty. The question in this trial was whether the accused knew (or should have known) the uses to which the gas would be put.

Applying the foreseeability standard, the manipulative view of Facebook's responsibility provides even greater conceptual leeway for attributing criminal responsibility to Meta than the neutrality view. Once we concede that Facebook creates a sphere of manipulation, which in turn is part of the story of toxic speech, then what the company needs to foresee is not (only) what users might do with the platform,[13] as would be the case under the neutrality view, but whether and to what extent its own algorithms are manipulative and how they facilitate the manipulation of "rogue" actors. The concern, put differently, is not (only) whether extremists are using the platform and how to muzzle them but whether the platform's recommendation algorithms are creating extremists and assisting them, and how to prevent this. The attribution of criminal responsibility no longer depends on the knowledge that the corporation has (or should have) of the deeds of some of its users, which would be a Herculean task given the size and diversity of Facebook's user base, but on the predictable impact of its business strategies and, more generally, its business model. Thus, even if Meta's intention in offering its platform is to obtain financial gain and not incite hatred, and even if the corporation has complied with due diligence to investigate the possible consequences of the use by third parties of the platform, this does not exempt it completely of its responsibility for the harmful effects of the content present on the platform (Hakim 2020).

Taking Meta to criminal court for its responsibility in fostering toxic speech will, of course, be an arduous, if not impossible, goal. The challenges are practical and normative, and they are multiple. Concerning the practical challenges, states in transitional societies can be expected to struggle financially and logistically to launch legal action against corporate actors who, by contrast to transitional states, may be wealthy and resourceful (but see Pereira 2022, who discusses some innovative ways in which Argentinian prosecutors brought corporate actors to justice). Action taken outside of transitional societies also has grim prospects. Global institutions such as the International Criminal Court are not set up to prosecute companies, and international criminal law generally does not allow for the prosecution of corporations. More importantly, the intent requirement, which I rejected earlier, is the relevant standard to attribute criminal responsibility for this court. Criminal prosecutions in places like the United States or the Europeam Union could have global ramifications. Yet matters are equally difficult here. Intermediary liability, as mentioned at the outset, shields platforms from criminal liability in many jurisdictions, most notably the United States.

There are also normative challenges to bringing Meta to court. I mention just one. It has been argued that eliminating intermediary liability may incentivize platform companies to conduct an overzealous content moderation policy to avoid any chance of facing liability, which in turn may lead them to indiscriminately target good as well as bad actors, or rather non-toxic as well as toxic speech. Instead of taking down toxic content from perpetrators, platforms may end up removing posts from, say, human rights activists denouncing atrocities. This is what happened when another platform—YouTube—removed videos from human rights organizations denouncing abuses during Syria's civil war.

This is a sensible point, but we should not draw the wrong inference from it. If intermediary liability is to be retained, it is because forcing companies to police speech in the platforms may turn out to be counterproductive, given current content moderation capabilities, not because platforms are not responsible for toxic speech. In other words, platform companies may be responsible for toxic speech in ways that would ordinarily make them criminally liable, but circumstances are such that we should hold them responsible in ways that do not involve criminal sanctions. The reasoning is analogous to the one Jeff Howard (2019) develops about the legal right of individuals to utter dangerous speech. In many jurisdictions, particularly in the United States, individuals are *legally* protected to spew toxic speech, even speech that incites violence. This is not because they have a *moral* right to free expression. Quite the contrary. They are under a moral duty not to express opinions that incite violence, a duty, moreover, that the state can coercively enforce. So where does the legal protection for toxic speakers come from? In Howard's account, it comes from the awareness of the detrimental consequences of state enforcement. Enforcement can be counterproductive, susceptible to political abuse, impossible to legally code without the risk of overinclusion, or it can deprive listeners of information that would allow them to exercise their intellectual virtues. For these reasons, it is best not to enforce the duty or, more radically, to extend a legal entitlement to people against its enforcement. This point reveals an inconsistency between certain justifiably created legal provisions and what Howards calls, following just war theory, deep morality. Extrapolating this argument, one could contend that intermediary liability is to corporations what the legal entitlement to spew toxic speech is to individuals. The basis for not criminalizing platforms' algorithmic manipulation is not that it is morally and criminally blameless: it is that it might be detrimental to do so.

This distinction is critical if we bear in mind that, as I mentioned earlier, the aim of criminal justice in transitional justice contexts is to publicly repudiate wrongdoing.

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

The expressive goals of punishment may be partially preserved if the decision not to pursue criminal prosecution against corporations like Facebook is widely understood to be part of an inevitable compromise between the repudiation of wrongdoing, on one hand, and the protection of digital platforms that *some* actors, such as human rights activists, may instrumentalize for good purposes, on the other.

Conclusion

In this paper, I argued that the growing importance of social media platforms in societies fractured by extraordinary conflicts poses challenges for transitional justice processes. The impact of these platforms in episodes of mass violence is relatively complex and still poorly understood. This complicates our understanding of the responsibility that platform companies bear for such violence. Focusing on the case of Meta, the purpose of this work was to contribute to delineating the conceptual and normative contours of such responsibility, taking seriously the role played by algorithmic influence within the platforms. I contend that transitional justice processes can only be conducted with adequate focus and depth when platforms are understood to be co-authors of toxic speech.

Acknowledgments

The author is indebted to audiences at the National Autonomous University of Mexico; the University of Edinburgh (in the Law School and the Centre for Ethics and Critical Thought); the Ethics and Public Policy Workshop, organized by Douglas MacKay and Christian Pérez Muñoz; and the 2024 APSA Conference. In particular, he would like to thank Michael Blake, Mihaela Mihai, Mathias Thaler, Tom Donahue-Ochoa, and the anonymous reviewers and editors of *Perspectives on Politics* for their valuable feedback. He is also grateful for the generous funding provided by project PAPIIT IA400523 and the APSA Summer Centennial Center Research Grant.

Notes

1 The company is under a legal obligation to enforce the Oversight Board's decision to keep or remove content but is not equally obligated to follow other recommendations. On the Oversight Board see Wong and Floridi (2023).

2 To be clear, what I call "neutrality" in this context is Meta's reluctance to remove a particular post from the platform, or to demote it, based on considerations other than its ability to "engage." The obvious exception is illegal content such as child pornography. The platform is neutral in the sense that whatever content produces engagement—humane or dehumanizing, antagonistic or conciliatory, crude or sophisticated, and so on—it will promote.

3 I have in mind, roughly, the distinction that many penal codes establish between different kinds of criminal intent (purposely, knowingly, recklessly, or negligently committing wrongdoing), from most to least serious in terms of degrees of culpability. These are legal distinctions, but they of course track moral intuitions.

4 This case was a precursor to two recent cases before the U.S. Supreme Court, *González v. Google* and *Twitter v. Taamneh.* The case discusses the corporation's legal, not moral, responsibility, for which "material contribution" is the legal test, but material contribution should not be equated to causal responsibility.

5 The corporation rebranded from Facebook to Meta in 2020, after this case.

6 I cannot discuss in detail the business model of companies like Meta. Suffice it to say that it is a model that depends on what Shoshana Zuboff (2019) calls the capture, analysis, and monetization of behavioral surplus, the essentials of "surveillance capitalism."

7 Anonymity is the affordance that most obviously encourages uninhibited speech and reduces incentives for self-control (Goldsmith and Wall 2022, 105).

8 Some scholars contend that platforms polarize and that antagonism between social groups on social media is associated with, and can statistically predict, violence when these groups meet in the real world (Gallacher, Heerdink, and Hewstone 2021; Munn 2020; Bail 2021; Karell et al. 2023; González-Bailón and Lelkes 2023). Conversely, others deny that users inhabit epistemic chambers altogether (Eady et al. 2019) or that adjusting algorithms (e.g., substituting algorithmic recommendation for reverse-chronological ordering of content) causes meaningful changes in political attitudes, such as ideological extremity, affective polarization, and even offline behavior (Guess et al., 2023a and 2023b; Nyhan et al. 2023). It should be noted, however, that "the effects of algorithms could be more pronounced in settings with fewer institutionalized protections (for example, a less-independent media or a weaker regulatory environment)" (Guess et al. 2023a). This is probably the case in a setting like Kenya and similar ones

9 This is a kind of technological harm that Wood (2021) calls generative harm, i.e., when technology induces harm on those who wield it, as when photo-sharing sites induce addiction in adolescents (see also Coeckelbergh 2022, 100-101, 108); generative harm is not about what users do with the technology (an instrumental harm) but about what the technology does to users, forming habits and desires.

10 In the social media literature, radicalization via recommender systems tends to refer to individuals who become exposed to extreme views and radicalized as a result. But "algorithmic" radicalization can also refer

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

to situations where individuals were radical to begin with, and as a consequence of it banded together more strongly or came to be assured of their views, taking them to be mainstream or at least legitimate.

11 Philosophical literature debates whether manipulation is inherently wrong or whether its wrongness depends on the context where it unfolds or the harm it produces. While I endorse the latter view, this is not the place to address the debate at length. See Jongepier and Klenk (2022), who discuss several ways to characterize manipulation based on its outcome, its process, or the set of norms it violates (e.g., a norm of proper influence).

12 Other important aspirations of transitional justice are reparations to victims and memorialization.

13 The neutrality view could be the basis for attributing criminal responsibility to Meta for tolerating the "weaponisation of social media" (Singh 2019) or for "criminal platforming" (Howard 2022). From this perspective, as already mentioned, Meta would have to show that it carried out human rights due diligence, inquiring into whether third parties could and would use the platform for wrongful purposes before making it available to them. In the case of Ethiopia, for instance, the inquiry would have been extraordinarily simple, since national and international human rights groups, including fact-checking organizations, approached the corporation to raise concerns about how the platform had been leveraged against ethnic minorities. Many of these companies complained that their concerns had been heard at best but not followed through (Jackson, Kassa, and Townsend 2022).

References

Alfano, Mark, J. Adam Carter, and Marc Cheong. 2018. "Technological Seduction and Self-Radicalization." *Journal of the American Philosophical Association* 4(3): 298–322.

Alfano, Mark, Amir Ebrahimi Fard, J. Adam Carter, Peter Clutton, and Colin Klein. 2020. "Technologically scaffolded atypical cognition: The case of YouTube's recommender system." *Synthese* (1-2):1–24.

Bail, Christopher. 2021. *Breaking the Social Media Prism: How to Make our Platforms Less Polarizing*. Princeton, NJ: Princeton University Press.

Benn, Claire, and Seth Lazar. 2022. "What's Wrong with Automated Influence." *Canadian Journal of Philosophy* 52 (1):125–48.

Bhargava, Vikram R., and Manuel Velasquez. 2021. "Ethics of the Attention Economy: The Problem of Social Media Addiction." *Business Ethics Quarterly* 31(3): 321–59.

Cassam, Quassim. 2022. *Extremism: A Philosophical Analysis*. Routledge: Abingdon.

Coeckelbergh, Mark. 2022. *The Political Philosophy of AI*. Polity: Cambridge.

Dangerous Speech Project. 2021. "Dangerous Speech: A Practical Guide," April 19 (https://dangerousspeech.org/guide/).

Eady, Gregory, Jonathan Nagler, Andy Guess, Jan Zilinsky, and Joshua A. Tucker. 2019. "How Many People Live in Political Bubbles on Social Media? Evidence from Linked Survey and Twitter Data." *SAGE Open* 9(1). http://doi.org/10.1177/2158244019832705

Elliot, Vittoria, and Dell Cameron. 2022. "A New Lawsuit Accuses Meta of Inflaming Civil War in Ethiopia." *Wired*, December 13 (https://www.wired.com/story/meta-hate-speech-lawsuit-ethiopia/).

Engelen, Bart, and Thomas Nys. 2022. "Commercial Online Choice Architecture: When Roads Are Paved with Bad Intentions." In *The Philosophy of Online Manipulation*, ed. Fleur Jongepier and Michael Klenk. New York: Routledge.

European Union. 2022. 2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), (https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065).

Forestal, Jennifer. 2022. *Designing for Democracy. How to Build Community in Digital Environments*. New York: Oxford University Press.

Gallacher, John. D., Marc. W. Heerdink, and Miles Hewstone. 2021. "Online Engagement Between Opposing Political Protest Groups via Social Media Is Linked to Physical Violence of Offline Encounters." *Social Media + Society* 7(1). http://doi.org/10.1177/2056305120984445

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Information Society and Culture Hidden Decisions that Shape Social Media*. New Haven, CT: Yale University Press.

Goldsmith, Andrew, and David S. Wall. 2022. "The Seductions of Cybercrime: Adolescence and the Thrills of Digital Transgression." *European Journal of Criminology* 19(1): 98–117.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7(1). http://doi.org/10.1177/2053951719897945

Groth, Olaf, Mark Nitzberg, and Stuart Russell. 2019. "AI Algorithms Need FDA-Style Drug Trials." *Wired*, August 15.

González-Bailón, Sandra, and Yphtach Lelkes. 2023. "Do Social Media Undermine Social Cohesion? A Critical Review." *Social Issues and Policy Review* (17): 155–180.

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

Guess, Andrew M. et al. 2023a. "How Do Social Media Feed Algorithms Affect Attitudes and Behavior in an Election Campaign?" *Science* 381:398–404.
——. et al. 2023b. "Reshares on social media amplify political news but do not detectably affect beliefs or opinions." *Science* 381: 404–408.
Gunn, Hanna. 2022. "Is There a Duty to Disclose Epistemic Risk?" In *The Philosophy of Online Manipulation*, ed. Fleur Jongepier and Michael Klenk. New York: Routledge.
Hakim, Neema. 2020. "How Social Media Companies Could Be Complicit in Incitement to Genocide," *Chicago Journal of International Law* (21)1: 83–117.
Hayner, Priscilla B. 2010. *Unspeakable Truths: Transitional Justice and the Challenge of Truth Commissions*. New York: Routledge.
Howard, Jeffrey. 2019. "Dangerous Speech." *Philosophy and Public Affairs* 47(2): 208–54.
——. 2022. "Extreme Speech, Democratic Deliberation, and Social Media." In *The Oxford Handbook of Digital Ethics*, ed. Carissa Veliz, 1–22. Oxford: Oxford University Press.
Jackson, Jasper, Lucy Kassa, and Mark Townsend. 2022. "Facebook 'Lets Vigilantes in Ethiopia Incite Ethnic Killing,'" *The Guardian*, February 20.
Jakobsen, Line Jespersgaard. 2023. "The Emerging Corporate Turn in Transitional Justice." *Cooperation and Conflict* 58(4): 561–71.
Jongepier, Fleur, and Michael Klenk. 2022. "Online Manipulation: Charting the Field." In *The Philosophy of Online Manipulation*, ed. Fleur Jongepier and Michael Klenk. New York: Routledge.
Karell, Daniel, Andrew Linke, Edward Holland, and Edward Hendrickson. 2023. "'Born for a Storm': Hard-Right Social Media and Civil Unrest." *American Sociological Review* 88(2): 322–49.
Keller, Daphne. 2021. "Amplification and Its Discontents." *Occasional Papers of the Knight First Amendment Institute*, Columbia University.
Lazar, Seth. 2023a. "Governing the Algorithmic City." *Tanner Lectures*, (https://www.youtube.com/watch?v=MzRWdpB39qw&t=397s).
——. 2023b. "Communicative Justice and the Distribution of Attention." *Tanner Lectures*, (https://www.youtube.com/watch?v=97U8BZAbJYo).
Lepora, Chiara, and Robert Goodin. 2013. *On Complicity and Compromise*. Oxford: Oxford University Press.
Massachi, Sahar. 2021. "How to Save our Social Media by Treating It Like a City." *MIT Technology Review*, December 20.
Maynard, Jonathan Leader. 2022. *Ideology and Mass Killing: The Radicalized Security Politics of Genocides and Deadly Atrocities*. Oxford: Oxford University Press.
Meaker, Morgan. 2022. "How the Kremlin Infiltrated Russia's Facebook." ()https://www.wired.co.uk/article/vk-russia-democracy.
Menn, Joseph. 2023. "Russians boasted that just 1% of fake social profiles are caught, leak shows." https://www.washingtonpost.com/technology/2023/04/16/russia-disinformation-discord-leaked-documents/?mc_cid=6cac3e76a5&mc_eid=9352836baf
Meta. 2021a. "An Update on Our Longstanding Work to Protect People in Ethiopia." https://about.fb.com/news/2021/11/update-on-ethiopia/
——. 2021b. "Corporate Human Rights Policy." https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf
Miller, Zinaida. 2008. "Effects of Invisibility: In Search of the 'Economic' in Transitional Justice," *International Journal of Transitional Justice* 2(3): 266–91.
Monsees, Linda, Tobias Liebetrau, Jonathan Luke Austin, Anna Leander, and Swati Srivastava. 2023. "Transversal Politics of Big Tech," *International Political Sociology* 17(1): 1–23.
Munn, Luke. 2020. "Angry by Design: Toxic Communication and Technical Architectures." *Humanities and Social Sciences Communications* (7)53. http://doi.org/10.1057/s41599-020-00550-7
Murphy, Collen. 2017. *The Conceptual Foundations of Transitional Justice*. New York: Cambridge University Press.
——. 2022. "Transitional Justice and Technology." *Social Philosophy and Policy* 38(2): 170–90.
Nguyen, C. Thi. 2020. "Echo Chambers and Epistemic Bubbles." *Episteme* 17(2): 141–61.
Nino, Carlos Santiago. 1997. *Juicio al mal absoluto*. Buenos Aires: Emecé.
Noggle, Robert. 2022. "The Ethics of Manipulation." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. (https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation/).
Nyhan, Brendan, Jaime Settle, and Emily Thorson. 2023. "Like-Minded Sources on Facebook Are Prevalent but Not Polarizing." *Nature* 620:137–44
Oversight Board. 2021. "Alleged Crimes in Raya Kobo." (https://oversightboard.com/decision/FB-MP4ZC4CC/).
——. 2022. "Tigray Communication Affairs Bureau," (https://oversightboard.com/decision/FB-E1154YLY/).
Payne, Leigh, Gabriel Pereira, and Laura Bernal-Bermúdez, eds. 2020. *Transitional Justice and Corporate Accountability from Below: Deploying Archimedes' Lever*. Cambridge: Cambridge University Press.
Pereira, Gabriel. 2022. "Corporate Accountability in Argentina: Fighting Corporate Impunity in Provincial Transitional Justice Context." In *Transitional Justice and Corporate Accountability from Below: Deploying*

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press

Archimedes' Lever, ed. Leigh Payne, Gabriel Pereira, and Laura Bernal-Bermúdez (eds.) (2020).
Reviglio, Urbano. 2017. "Serendipity by Design? How to Turn from Diversity Exposure to Diversity Experience to Face Filter Bubbles in Social Media." *Internet Science. Lecture Notes in Computer Science* 10673:281–300.
Richards, Imogen, and Mark Wood. 2020. "Legal and Security Frameworks for Responding to Online Violent Extremism: A Comparison of Far-Right and Jihadist Contexts." In *The Handbook of Collective Violence.*, ed. Carol A. Ireland, Michael Lewis, Anthony Lopez, and Jane L. Ireland. London: Routledge
Sahebi, Siavosh, and Paul Formosa. 2022. "Social Media and Its Negative Impacts on Autonomy." *Philosophy and Technology* 35(3): 1–24.
Simons, Joshua. 2023. *Algorithms for the People: Democracy in the Age of AI*. Princeton, NJ: Princeton University Press.
Singh, Shannon. 2019. "Move Fast and Break Societies: The Weaponization of Social Media and Options for Accountability under International Criminal Law." *Cambridge International Law Journal* 8:331–42
Srivastava, Swati. 2021. "Algorithmic Governance and the International Politics of Big Tech." *Perspectives on Politics* 21(3): 989–100.
Straus, Scott. 2012. "Retreating from the Brink: Theorizing Mass Violence and the Dynamics of Restraint." *Perspectives on Politics* 10(2): 343–62.
Stray, Jonathan, Ravi Iyer, and Helena Puig Larrauri. 2023. "The Algorithmic Management of Polarization and Violence on Social Media." *Knight First Amendment Institute* 23-05 (Aug. 22), (https://knightcolumbia.org/content/the-algorithmic-management-of-polarization-and-violence-on-social-media)
Sunstein, Cass. 2017. *#Republic: Divided Democracy in the Age of Social Media*. Princeton, NJ: Princeton University Press.
Susser, Daniel, Beate Roessler, and Helen Nissenbaum. 2019. "Online Manipulation: Hidden Influences in a Digital World." *Georgetown Law Technology Review* 4:1–45.
Swenson, Geoffrey, and Johannes Kniess. 2021."International Assistance after Conflict: Health, Transitional Justice and Opportunity Costs." *Third World Quarterly* 42(8): 1696–714.
Tirrell, Lynne. 2017. "Toxic Speech: Toward an Epidemiology of Discursive Harm." *Philosophical Topics* 45(2): 139–61.
United Nations. 2011. *Guiding Principles on Business and Human Rights.* (https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf).
Vaidhyanathan, Siva. 2018. *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. New York: Oxford University Press.
van der Merwe, Hugo, and M Brinton Lykes. 2022. "Transitional Justice and Corporate Accountability: Introducing New Players and New Theoretical Challenges." *International Journal of Transitional Justice* (16)3**:** 291–97.
van Vuuren, Hennie, and Michael Marchant. 2022. "Transitional Justice and Economic Crimes: Innovative Approaches from South Africa." In *Economic Actors and the Limits of Transitional Justice: Truth and Justice for business Complicity in Human Rights Violations*, ed. Leigh A. Payne, Gabriel Pereira, and Laura Bernal-Bermúdez. Oxford: Oxford University Press.
Wildman, Nathan, Natascha Rietdijk, and Alfred Archer. 2022. "Online Affective Manipulation." In *The Philosophy of Online Manipulation*, ed. Fleur Jongepier and Michael Klenk. New York: Routledge.
Williams, James. 2018. *Stand Out of Our Light: Freedom and Resistance in the Attention Economy*. Cambridge: Cambridge University Press.
Wong, David, and Luciano Floridi. 2023. "Meta's Oversight Board: A Review and Critical Assessment." *Minds & Machines* 33(2): 261–84.
Wood, Mark. 2017. "Antisocial Media and Algorithmic Deviancy Amplification: Analysing the id of Facebook's Technological Unconscious." *Theoretical Criminology* 21(2): 168–85.
——. 2021. "Rethinking how Technologies Harm." *The British Journal of Criminology.* 61(3):627–647.
Zelalem, Zecharias and Peter Guest. 2021. "Why Facebook keeps failing in Ethiopia," *Rest of the World*, November 13, https://restofworld.org/2021/why-facebook-keeps-failing-in-ethiopia/
Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

https://doi.org/10.1017/S1537592724001282 Published online by Cambridge University Press