

Reconstructing the Web: Towards a Diagnosis of the Moral Pathologies of Platforms¹

Abstract: Influential reconstructions of moral communication rarely take new forms of technological mediation into account. This paper does so, considering how the practical pressures that shape moral communication on social media differ from those that often shape it offline in order to develop a set of novel theoretical explanations of the moral pathologies of social media. First, because moral communication on social media is often at considerable remove from *action situations* and *concrete communities* and so from the need to coordinate successfully and maintain healthy relationships, it selects for moral communication that is more critical than coordinative, and criticism that is untempered. Second, because social media facilitates witnessing wrongs more than it does wrongdoing, it selects for expressions of indignation, but it also frustrates the conversational practices by which we work through the reactive attitudes. We should anticipate all this to have important consequences for moral language, understanding, and socialization.

¹ Special thanks to Jan-Werner Müller and Philip Pettit, both of whom read and commented on multiple early drafts, to Jeff Howard for incisive feedback at a much later stage that forced me to clarify a number of issues and significantly improved the paper, and to Daniel Wodak for invaluable framing advice. Thanks also to Chuck Beitz, Daniel Epstein, Andy Guess, Anna Jurkevics, Matthieu Queloz, Max Ridge, Helen Ross, Annie Stilz, Ophelia Vedder, and audiences at the 2022 Northwestern Graduate Conference and APSA 2023, for helpful written feedback, questions, and conversation, and to Miranda Fricker and Tori McGeer—the influence of whose written work will be obvious—for their guidance through the literature on moral address in a memorable graduate seminar at Princeton the fall of my first year. Comments from three anonymous referees at *Political Philosophy* and two at *European Journal of Political Theory* also improved the paper significantly.

Moral Talk Goes Online

Around a decade ago, prevailing sentiments about social media began to darken. Now, many agree that platforms, at least in their current form, are pathological, and their study is in a correspondingly *diagnostic* mood. Because some of the pathologies in question are new, diagnosis is partly a conceptual endeavor, involving naming and developing theoretical descriptions of the problems at hand. But historically, such diagnoses have also involved the development of theoretical *explanations*, actual or counterfactual, of the pathology diagnosed. Think, for an example of the latter, of Rousseau's genealogy of *amour-propre* from the dynamics of social evaluation in the second *Discourse*. There, Rousseau both names a social pathology that he sees around him—*amour-propre*, or the inflammation thereof—and explains how it could have emerged (Kolodny 2010; Neuhouser 2008).²

In what follows I employ a strategy broadly in the spirit of Rousseau's, focusing more narrowly on the social dynamics of *blame*, to develop a novel set of theoretical explanations of the moral pathologies of social media. The complex of pathologies— affective polarization (Bail, 2021; Iyengar et al., 2019; Iyengar and Westwood, 2015; Spinner-Halev and Theiss-Morse, 2024), moral outrage (Brady et al., 2021; Brady, Crockett, and Van Bavel, 2020; Nguyen and Williams, 2020), and public shaming (Billingham and Parr, 2020a and 2020b; Fritz 2021; Frye, 2021 and 2022), to list a few—is familiar, but their explanation is still nascent relative to that of epistemic pathologies like

² Other exemplary works in this diagnostic mode include Marx on alienation and Durkheim on *anomie*. For more general discussions of the diagnostic mode in philosophy, see Honneth (2007: ch. 1) and Neuhouser (2023). For more recent examples of such theoretical explanation in philosophy, albeit which aim to vindicate some practice or institution rather than diagnose pathological manifestations thereof, see Nozick (1974); Craig (1990); Pettit (2023) and other recent work; Williams (2002); see Queloz (2021) for discussion. Related, though not explicitly genealogical, approaches include rational-reconstructive explanations (Brennan et al., 2013: ch. 6) and paradigm-based explanations (Fricker, 2016), which I engage with closely in what follows.

misinformation. So while *echo chamber* is by now a household term,³ there remains a relative dearth of accounts of the roots of social media’s moral malaise.⁴ What, then, could help explain why people get so mad, and so moralistic, online?⁵ Why are the reactive attitudes and their expression so inflamed?

I have four connected hypotheses. We should, I argue, anticipate platforms—or, more precisely, patterns of mediation frequently, though not always or exclusively, instantiated thereby—to select for (1) *critical* rather than *coordinative* moral communication, and, relatedly, pathological forms of criticism, including (2) *untempered criticism* in general and (3) *indignant*, (4) *non-conversational* expressions of blame in particular. We should anticipate (1) and (2)—

³ Note the echo chamber hypothesis is much debated among philosophers (Nguyen, 2020; also see Dorst, 2020 and 2023) and social scientists (Bail et al., 2018; Guess, 2021; Prior, 2013).

⁴ But note the beginnings of scholarly discussion in political theory about what explains online public shaming or at least pathological manifestations thereof. Frye (2022: 123) explains public shaming in terms of the scale of social media, which allows many people to enforce norms at once, while Forestal (2024) argues that other factors bear on whether public shaming is pathological or productive, importantly the network structure in which it occurs. Productive public shaming, Forestal contends, is more likely in closed networks than open ones. My account identifies further properties of social media that might explain pathological public shaming, including depracticalization, depersonalization, and selection for indignation. Regarding whether these are properties of social media generally or of more particular network structures, the account I develop below suggests that, while depracticalization and depersonalization may be in some sense consequences of platform mediation per se, different networks will be more or less depracticalized and depersonalized. My hunch is that something similar is true of selection for indignation, depending on how much opportunity they provide for interpersonal wrongdoing (e.g., because they facilitate economic exchange) vs. the audience size they afford for different communicative interactions. More work is required to determine how my account interacts with Frye’s and Forestal’s, and the varying degrees to which the network properties we identify explain (pathological vs. productive) online communication is ultimately an empirical question. I look forward to seeing this conversation develop further. Many thanks to one of EJPT’s anonymous referees for directing me to Forestal’s recent paper.

⁵ The moral and epistemic pathologies of platforms may well be connected. For a while, it was not uncommon to see explanations of the moral pathologies of social media in terms of echo chambers. This would not survive the echo chamber hypothesis being debunked. More recently, some have suggested that outrage might rather explain the spread of misinformation (McLoughlin et al., 2024). If so, we’re wanting for explanations of outrage—hypotheses on which I offer here.

“untempered criticism,” let’s just say—because normative talk on platforms is often highly *depracticalized* and *depersonalized*, by which I mean, respectively, “used for purposes other than coordinating action”⁶ and “outside of relational contexts,” both of which would typically temper the way people criticize others. And we should anticipate (3) and (4) because platforms increase the number of wrongs that most people witness in a day more than they increase wronging, thereby selecting for indignation while simultaneously facilitating its public expression yet making the kinds of moral conversations through which we typically defuse it (and other reactive attitudes) more difficult. We should also, I further argue, expect (1)-(4) to have various and concerning downstream effects on moral language and socialization.⁷ Thus I also identify a novel set of metanormative concerns that bear on social media’s design related to the protection and cultivation of our moral-communicative practices, distinct from more familiar concerns like the protection of free speech or the social-epistemic consequences of platforms.

To generate these hypotheses, I begin from received theoretical models of moral communication and its role in human social life, which often already aim to explain the practice or particulars thereof in terms of its functional role therein (Fricker, 2016: 1-2), then consider how the introduction of different patterns of mediation would interact with these. The general idea is that these models typically suppose the communicative practices in question emerge from or occur in practical environments⁸ of certain kinds, and these play some role in explaining the

⁶ By “depracticalized” I don’t mean that they no longer pertain to the “ought” or the “should,” but rather that they’re un- or at least less directly connected to non-communicative forms of goal-directed action like work or other joint practical endeavors. Thus we could describe the problem as follows: lots of talk online is moralized without being practicalized—unconstrained, in other words, by practical necessity.

⁷ By “moral socialization” I mean: the social processes by which moral agency is developed and maintained.

⁸ Let a practice’s *practical environment* be the state of affairs that explain or shape that practice, including the practical pressures which bear on actors’ conduct and other features of that environment which interact with their attitudes and dispositions in a plausibly explanatory way.

practice, yet platforms constitute practical environments that differ considerably from those typically supposed. Relatively scant attention has been paid to how these models might change when the practical environments they suppose do (but see Goldberg, 2021; Schwartz, 2021, ch. 2).⁹ Closer attention, I show, has considerable payoff.

My argument proceeds as follows. In the next section I defend the untempered criticism hypotheses. First I introduce Habermas's account of the reproduction of a shared normative language from *The Theory of Communicative Action* (1984 and 1987), which, importantly, foregrounds the language's coordinative function, then compare this to normative language as typically used on Twitter/X (and other similar platforms, like Bluesky). Talk on these platforms, I argue, is relatively depracticalized and depersonalized, and this, I contend, should select for untempered criticism. From there I consider some potential consequences for normative languages, and also the degree to which other platforms and media systems depracticalize and depersonalize. Then I turn to a specific kind of moral criticism—blame and its expression—and argue that platforms select for expressions of indignation (hypothesis 3) but also frustrate healthy moral conversation (hypothesis 4). Here I mostly follow Fricker's (2016) account of healthy and pathological forms of blame. Finally, I spell out more carefully the potential consequences of the preceding discussion for the maintenance of moral understanding and healthy moral socialization.

Before that, a few provisos. First, platforms are obviously diverse. My discussion often foregrounds the type of platform most famously instantiated by Twitter/X, which I think

⁹ Goldberg writes: "In this paper I want to use a model of conversation from philosophy of language to diagnose some of the things that seem to be going wrong in our online 'conversations.' The model that I will introduce... employs some very idealized assumptions about the aim of conversations and the conditions under which they take place. As idealized and simple as it is, though, I think this model enables us to shed light on some of the challenges of online discourse, and it brings a fresh perspective on these. In fact, it might be because of the idealizations that the model sheds light on the challenges of our online conversations: many of the assumptions that the model needs to make either fail to hold in online conversations..." 177-78.

instantiate a pattern of mediation that selects for all of (1)-(4) to a high degree. While other platforms often instantiate the properties that would explain (1)-(4), they do not always instantiate all of them to the same degree. Similarly, other media technologies (like broadcast) share some, but rarely all, of the relevant properties. I will note when this is so. To be clear, I do not view any this as deeply problematic for my view. I am most fundamentally interested in shedding light on how different patterns of mediation might interact with our communicative practices as moral, political, and social theorists have reconstructed them, in hopes of identifying potential sources of media pathologies awareness of which might eventually bear on platform design.

Second, obviously, my goal is not to test my hypotheses empirically, but instead to introduce the relevant concepts and make what must remain a provisional, theoretical case for their plausibility.¹⁰ That said, this exercise may suggest avenues for empirical investigation,¹¹ and many of my arguments do square well with cutting-edge work in the social and cognitive sciences of the internet, on which I rely at various points throughout.

Third, to say that digital media might tend to “select for” certain modes of communication need not entail any kind of technological determinism. I do not understand “select for” to mean something like “make impossible for individuals to act or feel differently,” but rather to suggest

¹⁰ As Kitcher (2022: 148) puts it: “... working in the interstices of the fields of first-order inquiry, philosophy helps all those fields and helps humanity to take full advantage of their epistemic achievements. It does so by offering cognitive resources—not theories of everything... but concepts, questions, suggestive arguments, and analogies.”

¹¹ As Maxime Lepoutre writes (2021: 58): “[...] consider that when social scientists investigate the empirical consequences of public expressions of anger, they do not look randomly, but instead proceed with some theoretical hypotheses regarding what types of effects might reliably be related to anger. The task of developing such hypotheses is one that philosophers can fruitfully contribute to. By analysing emotions’ phenomenology, intentionality, and relation to epistemic values, philosophical accounts of emotions shed light on what kinds of consequences we can expect emotions to produce.” This kind of theory-forward approach may be well justified in light of the relative lack of good data about social media made available to researchers by tech companies—on which see Guess and Lyons (2020: 11-12).

general tendencies of different media given influential philosophical accounts of human moral psychology and practice.¹²

Fourth, there are other angles from which to approach these issues. Notably, I do not say much about political economy in general, or the profit motives that drive social media companies to design particularly attention-grabbing platforms in particular, here. This is an important part of the story of social media, which has been discussed at length by journalists and scholars across the humanities and social sciences (book-length treatments include Fisher 2022; Nielsen and Ganter 2022; Vaidhyanathan 2018; Webster 2014; Williams 2018; Wu 2016; for normative work on attention, see: Lazar 2024; Watzl 2022; also see Brady et al. 2023). And it is hard to imagine that profit-driven design choices of technology companies do not partially explain the pathologies of social media. But it may also be—so I argue here—that platforms themselves instantiate patterns of mediation with inherent pathologies or at least strong tendencies thereto. This includes their propensity to capture our attention by facilitating the spread of morally salient information, which technology companies have found various ways to monetize. If there is at least some truth to my account, it would seem the problem is at least partly with platforms per se, rather than just with particular, attention-grabbing features of their design. (Compare, for example, the socially pathological consequences of widespread sale of handguns in the United States. While we could imagine designing less pathological handguns, the handgun itself likely has pathological tendencies.) Now, my argument is of course perfectly compatible with an explanation of the proliferation of these patterns of mediation partly in terms of the profit motive—Silicon Valley

¹² Note: there's not a single selection mechanism that captures all of what I'm interested in the paper. In the first section I'm interested in how some of the forces that usually temper the way we criticize others are weakened by digital mediation. The suggestion is that we can sometimes felicitously explain behavior not just in terms of the incentives to act in certain ways, but also the ones that are saliently absent—like, for example, how we might explain the growing rowdiness of a grade-school classroom in terms of the teacher's absence. In the following sections, my focus is on how our reactive attitudes interact with platforms' technological affordances—"selection" in that case has to do with the kinds of interaction the technology supports (e.g., sharing information about wrongs more than wrongdoing).

developed a profitable product—but it is not my aim to develop this line here. Beyond helping explain why platforms engage people’s moral sentiments and so extract attention as powerfully as they do, my account also deepens our understanding of why this particular and fairly ubiquitous mode of attention extraction is just so socially damaging: because it engages and corrodes our moral sentiments and -communicative practices.¹³

Finally, though my account suggests reasons for deep concern with the digital mediation of our moral-responsibility practices, it does not rest on or yield condemnation of individual moral experience or behavior. Indignation is often an apt response to much of the information on platforms (see, e.g., Srinivasan, 2018), it’s natural to express it, and unsurprising that people express it as they do. The problem is rather that attitudes and communicative practices that are normally socially functional are poorly suited to patterns of mediation that platforms often instantiate (Brady et al., 2023 make a similar point). If anything, the account identifies a set of potential (if partial) excuses in ethical discussions of at least some genuine misbehavior on social media, along with an account of how genuinely justified indignation and expressions thereof can still run amok.

Depracticalization, Depersonalization, and Untempered Criticism

This paper does not aim to run the gambit of our moral-communicative practices; it focuses largely on moral address, specifically blame. That said, blame is made particularly salient by predictable consequences of platform mediation that are worth discussing to begin.

Much moral talk on social media is *depracticalized* and *depersonalized*—it occurs outside of, respectively, action situations and relational/community contexts—and so is at some remove from the practical pressures endemic to these, including those arising from the needs to coordinate successfully to achieve individual and shared ends and to maintain healthy

¹³ Thanks to an anonymous referee for pushing me to clarify this.

relationships. These properties, I contend, select for criticism in general and for untempered criticism in particular. While some recent theoretical discussion of social media has foregrounded new practical pressures that platforms *create*—game-like incentives (Nguyen, 2021) or the opportunity for certain kinds of attention or recognition (Ceilutka, 2023; also see Tosi and Warmke, 2016)—my primary focus in this section is on the unappreciated inversion: the practical pressures that are present in much face-to-face conversation but relatively absent from some platforms, which temper the way that people criticize others. Obviously, a general understanding of the consequences of platform mediation will require careful attention to both; I cannot claim complete comprehensiveness here.¹⁴

To begin, compare conversation on Twitter/X to an example borrowed, with considerable modification of my own, from *The Theory of Communicative Action*, which Habermas deploys to illustrate how actors rely on a shared *lifeworld*¹⁵ in communicatively coordinating their action, and how this communication in turn reproduces their shared lifeworld (Habermas, 1987: 121).

Suppose there is a new hire on a rural jobsite; he is from a left-leaning town with a fairly egalitarian work culture. Three mornings in a row, the foreman tells him to go grab coffee—this, in keeping with the local norm that errands go to the new guy. The new guy is not aware of this norm and so bristles.

¹⁴ Thanks to Jeff Howard for pushing me on this.

¹⁵ Prior to Habermas, *lifeworld* was a phenomenological concept—understood in Husserl, e.g., as the stock of “background” knowledge on which an individual actor relies to successfully navigate the physical world. This knowledge remains in the background in that it doesn’t need to be salient to be relied upon in successfully accomplishing our tasks—as when, e.g., I open the door and step into your office without first checking that there is a floor to step onto. Habermas redefines the lifeworld in language-philosophical (rather than phenomenological) terms as “a culturally transmitted and linguistically organized stock of interpretive patterns,” a *common linguistic stock* of “background knowledge” on which communicative actors rely to communicate felicitously with each other and so coordinate successfully.

Things might go different ways from here. Exit is probably off the table—the new guy needs a paycheck, so he cannot just quit—which leaves him Loyalty or Voice (see Hirschman, 1970). If the former, the new guy takes a deep breath and continues to comply with the foreman's orders until the next new hire. He may come to revise his expectations for the jobsite; because he put in his time, the next new guy should as well—fair's fair. If the latter, come fourth morning, the new guy refuses to grab coffee. If so, some work is needed to restore the normative consensus that makes possible relatively frictionless coordination, and which is, to use the Habermasian vocabulary, a constitutive part of the workers' shared lifeworld. The foreman might apologize or at least clarify that no slight was intended. Other coworkers might inform the new hire that everybody around here does this kind of thing when they are new—fair's fair. The new guy might apologize for stubbornly refusing to get coffee. He might go get it, or they might work some other arrangement, potentially establishing a new local norm thereby.¹⁶

In both cases, the practical pressures associated with continuing to work together successfully at least partially shape interaction, including by increasing the pressure to maintain normative consensus or restore it when it breaks down. In the former, these pressures compel the new guy to suck it up, continue to run errands for the time being, and maybe to ultimately revise his normative understandings vis-à-vis the norm of chores for new hires; here, shared normative understandings are reproduced through successful coordination. In the latter, these practical pressures motivate quick restoration of consensus needed to coordinate well or work harmoniously; here, the need to resume effective coordination and joint activity and to maintain healthy working relationships creates pressure to restore normative understanding. And though, in the second case, the new guy does in fact communicate blame (in the form of resentment) to

¹⁶ In Habermas's model, this working-out ideally occurs through a more demanding form of communicative action (i.e., with more demanding constitutive norms), which he terms "discourse." Participants reflect on their established norm of grunt work for new hires, considering alternative possibilities, and trying to reestablish consensus by convincing each other of their position on grunt work through the force of the better argument.

his foreman, these pressures also presumably *temper* how he communicates his blame, incentivize the foreman to apologize, and so forth; the need to resume coordination, etc., favors a reasonably healthy conversation about the issue that inspired resentment in the first place. Obviously not every case of blame in action contexts and concrete communities will be healthy or productive in this way. What is important is that the practical pressures associated with effective coordination (especially in the pursuit of shared ends) and the maintenance of relationships do seem to incentive that kind of engagement.

It is essential to see that normative talk in both cases is embedded in an action situation and a concrete community (here, of workers). Habermas (1987: 123) writes that “*this action situation* presents itself as a field of actual needs for mutual understanding and of actual options for action...,” thus “for those involved, the action situation is the center of the lifeworld.” He defines an action situation a “segment of lifeworld contexts of relevance” that is “articulated through goals and plans of action...” and that “presents itself as a field of actual needs for mutual understanding and of actual options for action...” (ibid.: 123). Normative talk is focused on immediate, common goals, used to coordinate or act jointly, and shaped by the need to maintain working relationships. And the reproduction of normative language, or the lifeworld, occurs *thereby*.

Compare this to Twitter/X, Bluesky, or Threads. Suppose Sam wakes up one morning, opens Twitter, and sees something he finds annoying or upsetting. Maybe it is a photograph of Nancy Pelosi and Chuck Schumer taking a knee in Kente cloth, a video of a Milwaukee cop harassing a Black man without cause, or a drag queen reading books to children. So he sends off a tweet making fun of Pelosi and Schumer, criticizing Milwaukee PD, or railing against “gender ideology.”

This fun-making or criticism is in each case normative. It might even be said to have some kind of distant goal in view—e.g., police reform. Nevertheless, Sam’s tweet is meaningfully outside of the kind of action/relational context that shaped normative talk in the jobsite case. It

is unrelated to his immediate goals, thus it is unburdened of many of the pressures that constrain goal-oriented, in-person communicative interaction. To employ another Habermasian distinction, social media frequently severs the connection between *instrumental* action, on the one hand, and *communicative* action, on the other.¹⁷ For this reason, it also selects for *signaling* rather than *navigational* beliefs—beliefs used to signal, say, group membership, rather than to navigate the world (Funkhouser, 2017).

To be sure, people sometimes use social media to coordinate—to get word out about a protest, for example. And action (beyond communicative) definitely occurs online; clothing is purchased, money is donated, etc., though it bears note that these transactions are still depersonalized. The internet in general is an incredible tool for sharing practically relevant *information*. All that said, lots of mediated normative communication has no immediate goal or aspirations to coordinate and occurs outside the context of relationships.

This raises a concern: normative talk on Twitter/X does not match the description of normative talk in the jobsite case. On Twitter, the so-called action situation cannot be the center of the lifeworld simply because *there is no action situation* remotely resembling the one Habermas describes. Relatively few common goals, plans, needs, etc. shape the communication, nor do relationships, be they working ones or ones that we value more for their own sake. Much talk on Twitter/X is thus depracticalized and depersonalized.

These properties (and their inverses) can interact. Imagine, for example, a lunchtime back-and-forth about the death penalty on the jobsite.¹⁸ This normative communication is not coordinative, but disagreement is nevertheless tempered by the need to maintain healthy relationships, and for this reason might never veer into outright criticism. So the lunchtime

¹⁷ Putting it loosely, these are actions undertaken from instrumental rationality, on the one hand, and action oriented towards achieving mutual understanding, on the other. See Habermas (1984: 273-95) for discussion.

¹⁸ Thanks to Jeff Howard for suggesting this case.

conversation is personalized but depracticalized. That said, it is conceivable that the primary reason the maintenance of healthy relationships matters on the jobsite is qua instrumental to efficient work. So we might read the lunchtime conversation as practicalized in a broader, more attenuated sense. Alternatively, imagine people who do not know each other at all needing to rapidly coordinate towards some goal—triage after a natural disaster—where practical pressures to maintain relationships do not even emerge as such. In that environment normative talk is highly practicalized but not personalized.

I suspect the coincidence of depracticalization and depersonalization is especially pathological, because together they select for untempered criticism to a high degree.¹⁹ And both are, I think, predictable consequences of platform mediation, not just Twitter/X. Platforms facilitate communication between distant others, making depersonalized normative talk much easier and thus likelier. Because platforms primarily allow people to interact communicatively, but not in other ways, normative talk is often depracticalized.

What consequences might we expect of a high degree of both depracticalization and depersonalization? If these do, as I argue, select for untempered criticism, potentially quite a lot. Languages are reproduced through their use; normative cultures are reproduced at least in part through the use of normative language, too. And we have good reason to think that how—and, more broadly, the conditions in which—some normative language is used, therefore reproduced, will bear significantly on the content of that language and the normative culture that that language encodes.²⁰

¹⁹ One might push back, here, that lots of (especially political) conversation on Twitter/X and other platforms is practicalized/coordinative in sense—people talking about who to vote for, etc. I’m not sure this is an apt description of most political discourse online, which seems more typically *reactive* to political news than explicitly coordinative. See Gerbaudo (2022).

²⁰ I mean “content” broadly to include not only words or concepts but norms governing their felicitous or even apt use, including the tone of their use, etc.

Normative languages reproduced via criticism rather than coordination and in conditions in which practical pressures arising from the need to coordinate successfully or maintain healthy relationships are less weighty or ubiquitous will end up having different content than languages reproduced coordinatively and in thicker relational contexts. For one thing, we use different words and concepts in and for different normative practices, and so when a normative language is reproduced through one practice P rather than another Q, the content of that language will drift accordingly; the words/concepts used for P will become more central, further, finer-grained P-related words/concepts will arise. So, by hypothesis, a normative language can *drift* depending on the contexts in which it is used and thus reproduced (Brady and Crockett, 2024 make a similar point). A normative language that we use primarily for coordination or within relationships we value will evolve accordingly. So will one that we do not use much for coordination but primarily for depersonalized criticism.²¹ In fact, untempered criticism may itself contribute to further critical drift. For, when A blames B in an untempered way, which B experiences as in some sense too strong or over the top, B may be liable to respond in kind rather than in a way conducive to productive conversation.²² Obviously, no purely coordinative or purely critical normative language exists—it is plausible that neither could exist—but it seems equally implausible that normative languages could not differ in, e.g., their relative coordinative vs. critical content.

A few caveats regarding the scope of this point. First, different patterns of mediation, thus different platforms, are depracticalized and/or depersonalized to different degrees. Person-to-person and group messaging are deeply embedded in our offline lives, typically used to coordinate or stay in touch with family, friends, coworkers, etc. So, messaging is neither depracticalized nor

²¹ Blaming is a form of criticizing, though it is doubtful that all criticism is blame. Aesthetic criticism, for instance, does not have to involve blame. For an argument for moral criticism that's more like aesthetic criticism, see Dover (2020).

²² Points like these are relatively familiar in the reactive attitude literature. See Fricker (2016: 11) for discussion of pathological blame.

depersonalized, though I would not necessarily call iMessage, WhatsApp, etc. “platforms,” as we typically mean something more specific by that term.

Communication on Facebook is more often depracticalized than messaging. But, since Facebook networks hew closer to offline ones than, e.g., Twitter/X and Reddit (people “friend” people they know) they may not be so deeply depersonalized. That said, Facebook still brings users frequently into communicative contact with people they do not have independent relationships with, especially in groups and comment sections.

Reddit, by contrast, is mostly anonymous, therefore highly depersonalized.²³ But many people go to Reddit for practical advice—the “r/[topic]” organizing structure seems to facilitate this distinctively well—meaning that talk on Reddit is often practicalized in one sense, though differently than on the jobsite. Reddit’s practices of advice-giving and -seeking do not involve people coordinating towards a common goal the achievement of which depends on how they communicate but usually involves advice regarding private ends—often the same ones (how to care for one’s ailing cactus), but not usually common ones (efficiently building a hospital). I am not sure the latter would have the same tempering effects. For reasons of space, I cannot dwell more over the differences among existing platforms (but see Forestal 2022 for a helpful discussion). All else being equal, my account predicts that environments in which communication is more typically practicalized and/or personalized will contain less untempered criticism than environments in which it is highly depracticalized and depersonalized. This well squares with what I take to be the widely held view that Twitter and its hatchlings (X, Bluesky) are particularly acrimonious.

Just as depracticalization and depersonalization is not equally spread across platforms, depracticalized, depersonalized communication is not wholly endemic to platforms, but may, in fact, be an emergent property of mediation in general. The public sphere itself, at least on the

²³ Anonymity poses further important issues that I do not have the space get into here.

idealized Habermasian model, is almost certainly depracticalized. Unlike normative talk on the jobsite, normative talk in the Enlightenment coffee house—or, for that matter, among consumers of news in the mass public—is not directly coordinative, but rather takes the state, political parties, etc. as distant objects of criticism (Habermas, 1989; cf. Koselleck, 1988). The Enlightenment did, quite famously, introduce critical drift in the normative language—perhaps, my account suggests, as the result of the introduction of depracticalization and depersonalization made possible by the introduction of new media and subsequent rise of the modern public sphere. Supposing platforms will depracticalize communication further, we might simply be due for another critical shift. At some point, though, critical drift may have diminishing, then negative, returns. We want our normative languages to conduce to coordination *and* criticism, and I think the discussion above shows how the need to coordinate, maintain relationships, etc., can shape the use of normative language in a way that yields positive benefits. In the grand scheme of things, the public sphere is a new phenomenon, and for all the modern optimism about it, it seems that its growth via new media will yield new understandings of its potential pathologies.

Indignant, Non-Conversational Blame

Insofar as normative talk on social media often occurs outside of action situations, relationships, and concrete communities, thus at some remove from the pressures that characterize these, we should expect languages reproduced on social media to drift in critical directions. I will now suggest other structural features of social media select for specific forms of criticism: *indignant* and *third-personal* expressions of blame.

To begin with some terminology, blame is a form of moral address. *Moral address* denotes the set of communicative practices by which we hold each other morally responsible (most notably through blame), take responsibility (e.g., apologize), forgive, etc. Though these practices have important cognitive dimensions, they are emotionally laden as well. Different modes of moral address are understood to correspond to quasi-cognitive but emotionally laden *reactive attitudes*

like resentment (which we experience when others wrong us), remorse, guilt, or shame (when we do wrong and blame ourselves), and indignation (when we witness a wrong) (Strawson, 1962). Moral address typically involves communicating these attitudes: if so-and-so wrongs me, I express resentment; this helps him see that he has wronged me, which, as long as he is of decent moral psychology, should inspire remorse, which he communicates in apologizing (Fricker, 2016). Though we can both blame in a dispassionate way and/or without expressing reactive attitudes to others, it is essential to see that, ordinarily, blame is *emotionally laden* and *paradigmatically communicative*.

Different patterns of digital mediation seem to select for specific moral experiences, thus for specific reactive attitudes and modes of address. For one thing, for most people, it is more common to *witness a wrong* than it is to *do wrong* or *be wronged* online. Digital mediation therefore selects for indignation more than resentment or guilt, and social media in particular, along with comment sections and the like, facilitates public expressions of that indignation. But social media also frustrates interpersonal conversation, selecting for expressions of blame (resentment or indignation) in the grammatical third person—as in “he did this” rather than “you did this.”²⁴ There is nothing wrong with indignation or third-personal address in particular cases, but the proliferation of the two is likely to have pernicious consequences that I discuss in the next section.²⁵

Selecting for expression of indignation:

²⁴ The distinction at play here is between the *grammatical form of blame* (second-or third-personal) and the *emotional kind of blame* (resentful, indignant, or guilty). One can communicate resentment (at being wronged) to a second person (as in, “I blame [am resentful towards] you for what you did to me”) or about a third (as in, “I blame [...] him for what he did to me”). So too indignation.

²⁵ This is just a loose schema, which is not meant to imply homogeneous moral experience; resentment might feel different to different people; some people might be less prone than others.

To begin, why would more wrongs be witnessed than committed or endured online? One reason, just mentioned, is that offline wrongs are sometimes recorded and, when recorded, frequently uploaded to platforms. Once online, these records rapidly spread, in part because morally salient information is particularly likely to draw attention and to be further shared (Brady, Crockett, and Van Bavel, 2020). And so there is a glut of wrongs to witness online; millions more people may witness a single wrong than are in any way involved in it.

An example from my own life: I made my Twitter account, long since deactivated, in June 2020, during the massive wave of Black Lives Matter protests across the world. That month, my feed was a deluge of videos of peaceful protesters being brutalized by police officers. I saw hundreds such videos that month, and I was not alone in this—many of the videos were quickly viewed by tens of millions of people; you probably saw them too. While morally salient news, of course, circulated long before the advent of digital mediation, that most people have the power to take and immediately upload video represents a massive scaling-up of our collective power to disseminate and consume morally salient information, which will inspire indignation in those who receive it.

By contrast, while it is definitely the case that people wrong others online, I doubt there is *equivalent* proliferation of wrongdoing as there is wrong witnessing. For one thing, the internet is a communicative setting, somewhat detached from the physical world and so from physical vulnerabilities. While people can wrong each other online by violating each other's privacy, hurling slurs, cheating on each other (albeit in a slightly unusual way), manipulating or threatening or, if sufficiently tech savvy, stealing from each other, people cannot beat up or shoot each other online.²⁶ Thus, the range of wrongs possible online is limited to what we might call *informational* and *communicative* ones.

²⁶ Compare Kant's "Doctrine of Right," which takes the fact of our embodiment as its starting point. See Ripstein (2009); Stilz (2009).

Now, informational and communicative wrongs can be horrifying.²⁷ I do not mean to discount this. It also seems likely that, beyond the informational and communicative wrongs we already recognize, moral understandings will evolve to fit our new digital environments—that we will develop a richer normative language to describe them than we currently have.²⁸ Maybe resentment will catch up with indignation once new concepts emerge to describe the informational and communicative wrongs digital mediation distinctively facilitates.

I think not so, for a few reasons. The first regards communicative wrongs. It is that platforms do not just increase people’s ability to witness wrongs that occur offline, but also the communicative wrongs that occur *on platforms*. Suppose, for example, that someone hurls a slur on X; many more people are likely to witness it than if it occurred offline, unrecorded, which again selects for indignation.²⁹ So even if moral understandings do evolve such that they make salient a wider range of communicative wrongs, it is just in the nature of platform mediation to select for indignation, because each instance of platform-mediated communicative wrongdoing is *simultaneously a record of that wrongdoing*, in principle witnessable by many others.

Now, this point does not seem to apply to those informational wrongs enabled by digital mediation that are done “in private,” so to speak—notably cases in which someone wrongfully views, rather than expresses, something (e.g., Frowe and Parry, 2019). Digital mediation has significantly increased humanity’s collective power to wrong each other in this way. But even here, I think there is reason to doubt an equivalent proliferation of resentment would follow.

²⁷ The past few years have made clear that children are especially vulnerable to these sorts of wrongs online. See, e.g., Winston (2024).

²⁸ See Eidelson (2023) for a relevant and especially thoughtful discussion of some related dynamics, also pathologies, of communicative respect.

²⁹ There is an important question about just who is wronged by slurs, microaggressions, etc.—only their target, or all the people in the group the slur demeans? Thanks to Daniel Wodak for pushing me on this.

Consider revenge porn. A person who intentionally views revenge porn for the purposes of gratification³⁰ wrongs the person depicted. And a single revenge porn video may be viewed millions of times. But for one thing, resentment is not particularly sensitive to these sorts of numbers. One would feel similarly resentful if one's ex uploaded such a video, regardless of whether it is viewed 50,000 or 250,000 times. For another thing, this, like most such privacy violations that digital mediation distinctively facilitates, is a case which *many people* do wrong to *one person*, which, if anything, is a recipe for a proliferation of guilt, not resentment, though this would require activity of conscience on the part of at least some significant number of the wrongdoers.³¹ So even granting that digital mediation does distinctively facilitate certain kinds of wrongdoing, we should not necessarily predict an equivalent proliferation of resentment to follow therefrom.

To be clear, none of this implies that *each individual* will experience more indignation than resentment by dint of existence of platforms; think of what a trans woman on X probably endures these days, or the kinds of resentment fostered in “incel” and other insular, grievance-oriented online subcultures.^{32,33} What I do think it suggests is that for *most people*, this will be the case.³⁴ Platform mediation selects for indignation rather than resentment at the social scale. And this, because of properties that at least arguably inhere in the form.

³⁰ Or, perhaps, for purposes other than content moderation.

³¹ The discussion of the scaffolding of moral agency below may suggest that, given the absence of direct moral address of wrongdoers in these cases, activity of conscience is less likely.

³² Obviously in the former case the resentment is justified while in the latter it is not.

³³ Incels represent an interesting case. In principle, my arguments apply here. But I think what such communities probably add is a pattern of strong *identification with* morally salient information shared, such that it inspires resentment as much as indignation or the two bleed together, along with cultures of expression and affirmation of resentment in particular. It is unsurprising that birds-of-a-feather effects would pull salient interests but also salient resentments together online. I cannot go into as much detail as I would like on this here. I address these dynamics in more detail in my “Parasocial Recognition,” ms. available upon request.

³⁴ Thanks to a referee for suggesting this clarification.

Some of what I have said so far applies to mass media, too; broadcast, for example, similarly increases the morally salient information individuals receive, thus it may increase the *experience* of indignation.^{35,36} This would imply that what I am describing is part of a longer historical process that predates platforms or digital mediation in particular, which may well be right. But what platforms distinctively add, in terms of their affordances, is for individuals to immediately *express* their indignation back into the platform itself, and for others to witness that expression, etc., etc. This has meaningful consequences. It might contribute to the development of indignation-heavy normative languages and cultures—recall the point about reproduction of normative language above. We could also imagine this reaching an inflection point, ultimately desensitizing people to both indignation-worthy information and others’ expressions of indignation. Both are concerning developments, and ones that may well be occurring (Lewis, 2024).

Third-personal address and frustrated conversation:

The internet is a medium of communication, however it is not equally conducive to all kinds of communication. While different platforms may differ as above, many are much better for transmitting information and facilitating expression than for conversation per se (Goldberg, 2021), certainly the kinds of moral conversation that surround healthy, productive moral address. Indeed, social media commonly frustrates moral address and stifles moral conversation, in a few different ways.

One is that it allows many people to witness wrongs, or actions they take to be wrong, without being able to converse (productively) with either the wrongdoers or those who disagree with their moral judgments. This might be for a variety of reasons. Maybe the wrongdoers are not

³⁵ This of course depends on the degree to which the broadcasting entity is willing to show indignation-worthy information.

³⁶ Thanks to a referee for pushing me on this.

online or on the relevant platform. Even if they are on that platform, the sheer number of others blaming them can preclude their attending, let alone replying, to any utterance. There may also be strategic reasons to avoid replying at all to a deluge of digital blame. Alternatively, the wrong in question may have occurred years ago.

For these reasons, the kind of moral address that typically occurs on social media differs meaningfully from the kind that Miranda Fricker argues is paradigmatic and indeed healthy: second-personal conversation. In “What Is the Point of Blame?” she describes such a conversation proceeding roughly as follows: When A wrongs B, B feels resentful and communicates this resentment to A in a communicative act of blame, the illocutionary point of which is to inspire remorse in A. Things going well, A feels remorseful and communicates this to B with an apology. This process increases alignment between A and B’s moral understandings, thus helping them restore a common moral understanding.³⁷ Importantly, this kind of conversation often helps *resolve* the feeling of resentment—generally such conversation is an essential mechanism for working through our moral emotions. I detail Fricker’s account more below. But first, notice again how different things look on social media.

Suppose a video of horrible police brutality begins to circulate. Rightminded people feel indignant. But unlike the case in which one endures or witnesses a wrong offline and can express resentment or indignation directly at the person responsible, in this case the target(s) of the attitude will almost certainly be too distant for most to successfully communicate their reproach. Supposing the cop is on Twitter/X, there will be no point in tweeting at him. There is a good chance he will not see the tweet; if he does, it is unlikely the moral address it contains will receive uptake; it is almost certain he will not reply. It is not only that the cop is receiving a deluge of indignation, but also that these conditions will likely give him strong reasons to view and respond

³⁷ Notice the similarities to Habermasian discourse (fn. 16).

to the situation prudentially. His lawyers might advise against admission of guilt implied by public apology; beyond avoiding jailtime, he would be concerned to keep his job.

So, firing off my indignant tweet, I join a *chorus* of blamers. We might like or share each other's tweets and end up talking amongst ourselves about the incident or the state of American society. In this way, moral address on platforms is often more *choral* than *conversational*—or, if there is conversation, it is typically internal to the chorus and *third-personal*, about some third person who committed a wrong that many others have witnessed.

Of course, my chorus might confront another, defending the offending party. In that case, our feelings might be of some odd admixture of resentment and indignation—whatever it is that we feel when someone with whom we really are conversing simply denies that we should be feeling indignant at all about some matter we take to be morally serious. Here the publicity of conversation on some platforms is highly relevant. It allows any user of the platform to witness or even enter into the conversations of others, which can induce further moral emotion.

There are further conversational frustrations, too. Suppose, differently, you are scrolling through a feed and land on a recording of the Fuhrman tapes,³⁸ which you have never heard before. You might well feel similarly indignant. But here, moral address is also *temporally* frustrated; the Fuhrman tapes were recorded decades ago. While this temporal distance makes moral address much less apt or productive, this fact spares little indignation.³⁹ So here is another source of

³⁸ The Fuhrman tapes were a central piece of evidence in the OJ Simpson murder trial, supporting the defense's case that Simpson had been framed. The first police officer who arrived at the murder scene, Mark Fuhrman, had been recorded years prior using racial and homophobic slurs, bragging about brutalizing Black folks, planting evidence, and so on, in support his view that women were not cut out for police work. I am not going to link directly to the tapes, but for contemporaneous news coverage, see Claiborne and Wexler (1995).

³⁹ I, for one, had known about the Fuhrman tapes long before actually hearing them, but the mere knowledge of them did not make me indignant in the way that hearing those sentiments in Fuhrman's voice did.

frustration: we encounter things that make us rightfully indignant, in their full color, but which are for temporal reasons inapt for moral conversation.

I use “frustration” here not just signal that social media makes the normal conversations associated with moral address more difficult, but also because the experience of being unable to felicitously communicate the reactive attitudes is literally frustrating (see Funk, McGeer, and Gollwitzer, 2014). We have strong emotional reactions to being wronged and witnessing wrongs. Moral address in a conversational mode is satisfying because one often needs to hash-out one’s reactions to move productively beyond them. Absent such conversation, resentment can blister into *ressentiment* or indignation into its equivalently pathological form. So beyond getting a rise out of us by engaging our reactive attitudes, by frustrating the conversational modes that attach to them, social media also suspends us in that rise.

Moral Understanding and Moral Agency

All of this has consequences. Here the moral-address literature lends further insight. As noted above, Fricker (2016) argues that blame has a paradigmatic mode—resentful blame in its second-personal conversational form (as in “you wronged me”). Fricker terms this paradigmatic mode *Communicative Blame*. This mode is paradigmatic in the sense that it explains both the other forms of blame—guilt and indignation—and accounts for what we consider pathological manifestations of blame generally (ibid.: 1-4).

On Fricker’s model, Communicative Blame is a speech act, the illocutionary point of which is “to inspire that admixture of judgment and moral emotion that is *remorse*. It aims to bring the wrongdoer to see or fully acknowledge the moral significance of what they have done or failed to do” (ibid.: 8-9). Remorse is defined as “a cognitively charged moral emotion—a moral perception that delivers a pained understanding of the moral wrong we have done.” And blame and remorse are understood as “partner emotional cognitions, each bearing the same moral content (X wronged Y) but where the content is grasped from the opposite points of view of

wronged and wrongdoer... By way of this moral psychological calibration,” Fricker continues, “these two essential moral emotions—first blame and then remorse—work together to bring *increased alignment between the moral understandings of blamer and wrongdoer*,” which “will properly be achieved dialogically” (ibid.: 9). The blamer’s address constitutes the initial attempt at increased alignment—to use Fricker’s terminology, “jump-starting” the moral understanding of the blamed. Of course, the blamed might not be moved—they might really think they have done nothing wrong—and issue a rejoinder rather than an apology. This, in turn, might prompt the blamer to reconsider their stance. As long as one party successfully moves the other, moral understanding will be better aligned in both cases. “We may comprehend the point of blame through the lens of an ideal of converging moral understandings, while acknowledging of course that moral conflict is often too difficult and fraught for this kind of convergence to be possible” (ibid.). In its proper form, this kind of dialogue counts as “reasoning with emotional force” (ibid.).

I might add that even when neither party budges, moral dialogue can accomplish further good. For there is a sort of conversation through which people feel they have had the chance to make their appeal—they have been *heard* or *taken seriously* or *heeded*—even if they ultimately (as we often put it) “agree to disagree.” We have various strategies for achieving this slightly different resolution; we might avoid talking in an explicitly moral mode, instead simply expressing mutual concern about each other’s feelings—“while I don’t agree with you, it matters to me that I made you feel this way”—and even achieve some kind of understanding about how each of us will act in the future, if not alignment of moral understanding *per se*.⁴⁰ This just to say, moral dialogue can be productive in a slightly different way, too.

Supposing paradigmatic blame is as described, then, Fricker asks, what are we to make of the cases of distant, sometimes even *dispassionate* blame, which never gets communicated? She

⁴⁰ It is possible that Fricker intends to include this different case; whether it is included in her account will turn on the definitions of alignment and moral understanding.

notes that as moral and social relations become more complex, we should expect moral/social practices to emerge that are derivative of the paradigm.

Regarding the social side of things, there are clearly more kinds of relationship for the moral subject to stand in than the second-personal, so that reactions of blame will naturally extend themselves to apply in third-personal cases too, both nearby and very distant. But when we express blame at a distance without any second-personal communicative purpose, as we do for instance in the example of the impending French lorry driver strike, all the distinctive elements of the transformative purpose I have been at pains to attribute to Communicative Blame have fallen away—neither the second-personal perspective nor the aim to inspire remorse through the emotionally charged accusation of fault is any longer present: In whom am I trying to inspire remorse as miles from the scene I chat casually about which party I blame? No one. With whom am I upset? No one. Whom am I addressing? Not the wrongdoer, for sure. (Ibid.: 13)

Nonetheless, argues Fricker, we can detect a “residue” of Communicative Blame in even the least passionate instances of indirect blame. This residue is apparent in a few aspects of this kind of blame—most notably for our purposes in the conversational working-out and attempted alignment of our moral understanding of, e.g., the strike, with an antistrike friend, classmate, etc.

It is clear that Fricker is painting a slightly different picture of third-personal moral address than I do above. I think this is partly for dialectical reasons: the case in which address is not just third-personal and distant *but also dispassionate* is hardest for her because it is, on the face of it, furthest from Communicative Blame. For my own argumentative purposes, it is worth noting that not all such blame is dispassionate—indeed, that relics like the Fuhrman tapes, or maybe impassioned testimony from the striking drivers, do stir up more passionate blame in the indignant key. And this kind of blame does not conduce so well to aligning moral understanding. If we disagree about the strike, it is probably better for us to check our indignation at the conversation’s door. Moreover, though moral address on social media can be dialogical, it is often not so. Fricker envisions a conversation with an antistrike *friend*, not with strangers online. The conversation will occur against a backdrop of mutual understanding of the kind needed for, arguably constitutive of, a friendship. The need to protect that friendship (to argue gently enough

to do so, however gently that is) will also shape the conversation; recall the discussion of depracticalization above. Even good friends who argue ferociously trust each other in this and know when to let up. On social media, by contrast, one might find oneself in one chorus of blamers, dueling another. Teammates will assure each other of the team's overall correctness. When arguing with opponents, they may be more interested in scoring points with their teammates (e.g., getting likes), not reaching understanding (Nguyen, 2021). The most impassioned expressions will tend to attract the most attention (Brady, Crockett, and Van Bavel, 2020), scoring the most points. In these ways, conversation on social media differs from the kind of conversation Fricker describes.

So, to pull some strings together, while the kind of conversation that Fricker describes might sometimes be depracticalized (as when about the strike in France), it is also personalized and relatively dispassionate. But what seems potentially most pathological is not depracticalized, depersonalized moral communication, but depracticalized, depersonalized moral communication that is also *laden with moral emotion*. The reason is that the pressures associated with practicalized, personalized contexts—coordinating successfully, maintaining healthy relationships—temper the expression of moral emotion as described. While face-to-face moral address is often emotionally charged, it often occurs in an action situation and a relational/community context. And while the kinds of critical conversations that we associate with the public sphere (rightly or wrongly) or a philosophy seminar are often depracticalized and sometimes depersonalized, they are typically relatively dispassionate. Social media, by contrast, powerfully incites the reactive attitudes—especially, I have argued, indignation—but without these tempering elements present.

Important questions arise through this comparison of Fricker's picture of third-personal, dispassionate blame to blame on social media. One has to do with the illocutionary point of the latter. Does the blame I describe, for which social media selects, really aim to inspire remorse? If so, my account suggests that aim is often frustrated. Platforms appear to undermine people's ability to inspire remorse in others and communicate the remorse they feel. If so, we should

anticipate them to frustrate one key process by which we restore shared moral understandings or reaffirm them when they come into question.

Moral understanding is not all that is at stake. Moral agency itself might be. At least, any account on which the development or maintenance of moral agency involves participation in healthy moral-communicative practices will advise concern at media that undermine (the healthy manifestations of) these practices. There are many accounts we could draw from here (e.g., Habermas, 1990; Honneth, 1995), but for my purposes we can stay within the Strawsonian literature.

In her paper on blame, Fricker characterizes moral address as a *proleptic* mechanism; it involves treating the blamed *as if* they recognize the reasons they had against acting wrongly even if they did not in fact recognize those reasons. The blamer's treating the blamed this way can help the blamed appreciate the reasons they had missed. In "Scaffolding Agency," Victoria McGeer (2019) develops a similar thought further, arguing that, thanks to this proleptic mechanism, blame plays "a crucial role in developing and sustaining our capacity to recognize and respond to moral reasons." She defends this by developing an account of moral agency as a skill that is to some degree socially dependent or *scaffolded*. Moral agency is thus understood as accordion-like, expanding or contracting in different moral-social environments. And healthy moral-address practices first and foremost are what expand this capacity. "Your blame," McGeer writes,

calls for an appropriate reactive response from me—for example, recognition of my moral blunder. Presuming that I have no excuse for my bad behavior, such recognition involves granting that there were moral reasons for behaving otherwise to which I was not sensitive. But your blame demands that I go beyond mere recognition of this fact; it demands that I regret it, that I feel sorry about it, and that I am consequently motivated to take responsibility for doing better in the future, thereby *expanding my capacity* for tracking and responding to the reasons that there are. And I show all this by way of my reactive responses to your reactive blame—that is, by way of my guilt, shame, remorse, and contrition. (Ibid.: 314)

And so, the significance of our practices of moral address, indeed of the reactive moral emotions (resentment, indignation, remorse) themselves,

lies... in their powerful forward-looking character. For... they play an active role in shaping and regulating our future attitudes and behavior by way of developing and sustaining our capacity for (socially negotiated) moral-reasons-responsiveness. Hence, they play a central and critical role in scaffolding what it takes to be a morally responsible agent. (Ibid.)

If so, then if we predict social media to put strain on our moral emotions and practices of moral address, this is cause for deep concern. Blame has both healthy (moral-understanding-aligning, moral-agency scaffolding) and pathological forms. Insofar as social media selects for pathological forms of blaming, it may well undermine moral agency. Now, fortunately, most people do not spend most of their time on social media.⁴¹ And it is not as if healthy blame is *impossible* on social media. Still, the quality of moral communication on social media is independently important, and it would be surprising if there were no spillover effects to offline life.

The fundamental takeaways are as follows. The first is that designing social media or the internet in general better is an important sense a question of how its design positions it in relation to the offline spheres—to existing relationships and goal-oriented endeavors like work. Social media may not be, as it were, internally perfectible. We can design *against* the tendencies I describe, but I am less certain we can design them *away*, at least insofar as some of them arise from the basic affordances of social media rather than contingent features of particular platform designs. And our practices may also need to evolve to keep up with our changing practical environments. If not, platforms may continue to corrode our moral communication.

In a way, the concerns developed here are conservative in the literal sense: focused on the protection and cultivation of our normative culture, wary of new forces that seem to be threatening it. Given that we now have to make design choices about how our technology bears

⁴¹ That said, American teens spend an average of nearly 5 hours on social media per day (Rothwell, 2023).

on that culture, this kind of thinking seems to be necessary. That said, these kinds of quasi-conservative normative considerations are not the only ones, and sorting out how they weigh against others is an important further task.

References

- Bail CA (2021) *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton: Princeton University Press.
- Bail CA, Argyle LP, Brown TW, et al. (2018) Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of the Sciences* 115(37): 9216–9221.
- Billingham P and Parr T (2020a) Enforcing social norms: The morality of public shaming. *European Journal of Philosophy* 28(4): 997–1016.
- Billingham P and Parr T (2020b) Online public shaming: Virtues and vices. *Journal of Social Philosophy* 51(3): 371–390.
- Brady WJ and Crockett MJ (2024) Norm psychology in the digital age: How social media shapes the cultural evolution of normativity. *Perspectives on Psychological Science* 19(1): 62–64.
- Brady WJ, Crockett MJ and Van Bavel JJ (2020) The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science* 15(4): 978–1010.
- Brady WJ, Jackson JC, Lindström B, et al. (2023) Algorithm-mediated social learning in online social networks. *Trends in Cognitive Science* 27(10): 947–960.
- Brady WJ, McLoughlin K, Doan TN, et al. (2021) How social learning amplifies moral outrage expression in online social networks. *Science Advances* 7(33): 1–14.
- Brennan G, Eriksson L, Goodin RE, et al. (2013) *Explaining Norms*. Oxford: Oxford University Press.
- Ceilutka K (2023) The discontents of competition for recognition on social media: Perfectionism, resentment, and collective narcissism. *Philosophy & Social Criticism* 49(4): 409–430.
- Claiborne W and Wexler S (1995) Tapes hit home for L.A. blacks. *Washington Post*, 31 August. Craig E (1990) *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*. Oxford: Oxford University Press.
- Dorst K (2020) Reasonably polarized: why politics is more rational than you think. *Stranger Apologies*. https://www.kevindorst.com/stranger_apologies/rp

- Dorst K (2023) Rational polarization. *Philosophical Review* 132(3): 355–458.
- Dover D (2020) Criticism as conversation. *Philosophical Perspectives* 33(1): 26–61.
- Eidelson B (2023) The etiquette of equality. *Philosophy & Public Affairs* 51(2): 97–139.
- Fisher M (2022) *The Chaos Machine*. New York: Little, Brown and Company.
- Forestal J (2022) *Designing for Democracy: How to Build Community in Digital Environments*. Oxford: Oxford University Press.
- Forestal J (2024) Social media, social control, and the politics of public shaming. *American Political Science Review* 118(4): 1704–1718.
- Fricker M (2016) What's the point of blame? A paradigm based explanation. *Noûs* 50(1): 165–183.
- Fritz J (2021) Online shaming and the ethics of public disapproval. *Journal of Applied Philosophy* 38(4): 686–701.
- Frowe H and Parry J (2019) Wrongful observation. *Philosophy & Public Affairs* 47(1): 104–137.
- Frye H (2021) The technology of public shaming. *Social Philosophy and Policy* 38(2): 128–145.
- Frye H (2022) The problem of public shaming. *Journal of Political Philosophy* 30(2): 188–208.
- Funk F, McGeer V and Gollwitzer M (2014) Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin* 40(8): 986–997.
- Funkhouser E (2017) Beliefs as signals: A new function for belief. *Philosophical Psychology* 30(6): 809–831.
- Gerbaudo P (2022) Theorizing reactive democracy: The social media public sphere, online crowds and the plebiscitary logic of online reactions. *Democratic Theory* 9(2): 120–138.
- Goldberg SC (2021) The promise and pitfalls of online 'conversations.' *Royal Institute of Philosophy Supplement* 89: 177–193.
- Guess AM (2021) (Almost) everything in moderation: New evidence on Americans' online media diets. *American Journal of Political Science* 65(4): 1007–1022.

- Guess AM and Lyons BA (2020) Misinformation, disinformation, and online propaganda. In: Persily N and Tucker JA (eds) *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge: Cambridge University Press, 1–9.
- Habermas J (1984) *The Theory of Communicative Action*, Vol. 1. Boston: Beacon Press.
- Habermas J (1987) *The Theory of Communicative Action*, Vol. 2. Boston: Beacon Press.
- Habermas J (1989) *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Cambridge: MIT Press.
- Habermas J (1990) Individuation through socialization: On George Herbert Mead’s theory of subjectivity. In: *Postmetaphysical Thinking: Philosophical Essays*. Cambridge: MIT Press, 149–204.
- Hirschman AO (1970) *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Cambridge: Harvard University Press.
- Honneth A (1995) *The Struggle for Recognition: The Moral Grammar of Social Conflicts*. Cambridge: MIT Press.
- Honneth A (2007) *Disrespect: The Normative Foundations of Critical Theory*. Cambridge: Polity.
- Iyengar S, Lelkes Y, Levendusky M, et al. (2019) The origins and consequences of affective polarization in the United States. *Annual Review of Political Science* 22: 129–146.
- Iyengar S and Westwood SJ (2015) Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science* 59(3): 690–707.
- Kitcher P (2022) *What’s the use of philosophy?* Oxford University Press.
- Kolodny N (2010) The explanation of amour-propre. *Philosophical Review* 119(2): 165–200.
- Koselleck R (1988) *Critique and Crisis: Enlightenment and the Pathogenesis of Modern Society*. Cambridge: MIT Press.
- Lazar S (2024) Lecture II: Communicative justice and the distribution of attention. arXiv preprint arXiv:2410.20718. <https://arxiv.org/abs/2410.20718>

- Lepoutre M (2021) *Democratic Speech in Divided Times*. Oxford: Oxford University Press.
- Lewis T (2024) Outrage fatigue is real. These tips may help (interview with William Brady). *Scientific American*. 30 December.
- McGeer V (2019) Scaffolding agency: A proleptic account of the reactive attitudes. *European Journal of Philosophy* 27: 301–323.
- McLoughlin KL, Brady WJ, Okuhara AN, et al. (2024) Misinformation exploits outrage to spread online. *Science* 386(6725): 991–996.
- Neuhouser F (2008) *Rousseau's Theodicy of Self-Love: Evil, Rationality, and the Drive for Recognition*. Oxford: Oxford University Press.
- Neuhouser F (2023) *Diagnosing Social Pathology: Rousseau, Hegel, Marx, and Durkheim*. Cambridge: Cambridge University Press.
- Nguyen CT (2020) Echo chambers and epistemic bubbles. *Episteme* 17(2): 141–161.
- Nguyen CT (2021) How Twitter gamifies communication. In: Lacky J (eds) *Applied Epistemology*. Oxford: Oxford University Press, 410–436.
- Nguyen CT and Williams B (2020) Moral outrage porn. *Journal of Ethics and Social Philosophy* 18(2): 147–172.
- Nielsen RK and Ganter SA (2022) *The Power of Platforms: Shaping Media and Society*. Oxford: Oxford University Press.
- Nozick R (1974) *Anarchy, State, and Utopia*. New York: Basic Books.
- Pettit P (2023) *The State*. Princeton: Princeton University Press.
- Prior M (2013) Media and political polarization. *Annual Review of Political Science* 16: 101–127.
- Queloz M (2021) *Genealogy as Conceptual Reverse-Engineering*. Oxford: Oxford University Press.
- Ripstein A (2009) *Force and Freedom: Kant's Legal and Political Philosophy*. Cambridge: Harvard University Press.

Rothwell J (2023) Teens spend average of 4.8 hours on social media per day. Gallup. 13 October. <https://news.gallup.com/poll/512576/teens-spend-average-hours-social-media-per-day.aspx>

Schwartz O (2021) *Sociological Theory for Digital Society: The Codes that Bind Us*. Cambridge: Polity Press.

Spinner-Halev J and Theiss-Morse E (2024) *Respect and Loathing in American Democracy: Polarization, Moralization, and the Undermining of Equality*. Chicago: University of Chicago Press.

Srinivasan A (2018) The aptness of anger. *Journal of Political Philosophy* 26(2): 123–144.

Stilz A (2009) *Liberal Loyalty: Freedom, Obligation, and the State*. Princeton: Princeton University Press.

Strawson PF (1962) Freedom and resentment. *Proceedings of the Aristotelian Society* 62: 1–25.

Tosi J and Warmke B (2016) Moral grandstanding. *Philosophy & Public Affairs* 44(3): 197–217.

Vaidhyanathan S (2018) *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. Oxford: Oxford University Press.

Watzl S (2022) The ethics of attention: An argument and a framework. In: Archer S (eds) *Salience: A Philosophical Inquiry*. New York: Routledge, 89–112.

Webster JG (2014) *The Marketplace of Attention: How Audiences Take Shape in a Digital Age*. Chicago: University of Chicago Press.

Williams B (2002) *Truth and Truthfulness: An Essay in Genealogy*. Princeton: Princeton University Press.

Williams J (2018) *Stand Out of Our Light: Freedom and Resistance in the Attention Economy*. Cambridge: Cambridge University Press.

Winston A (2024) There are dark corners of the internet. Then there's 764. WIRED. 13 March. Available at: <https://www.wired.com/story/764-com-child-predator-network/>

Wu T (2016) *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. New York: Alfred A. Knopf.