CONTENT MODERATION ONLINE: REGULATION *EX ANTE* VERSUS *EX POST*

*Vincent Chiao**
*Alon Harel***

Recent years have witnessed proliferating calls for technology and social media companies to more aggressively police the speech of their users. Social media companies have drawn criticism both for being too aggressive and too lax in censoring their users' speech. While this controversy is typically framed in terms of the extent and significance of principles of free speech, we re-frame the debate in terms of a contrast between ex ante prevention and ex post punishment. Ex ante prevention operates as a form of censorship, preventing objectionable speech from occurring, whereas ex post punishment operates by censuring objectionable behavior after it has materialized. Content moderation operates in an ex ante manner by preventing targeted speech from reaching an audience, whereas ex post remedies are more diffuse, ranging from informal disavowals and condemnatory statements by other users or the platform itself to "de-platforming" offenders to, in extreme cases, formal legal actions.

We identify four factors that bear on the choice between ex ante prevention and ex post punishment of online speech. These are the closeness or fit between a substantive type of wrong and its codification in a rule; the potentially asymmetric costs of precaution; deliberative transparency; and the value of normative adaptation. Since these factors do not necessarily point in the same direction, the choice between ex ante prevention and ex post punishment of speech requires a substantive value judgment. That said, we argue that, in general, ex ante prevention is most appropriate when applied to tightly specified, high-stakes expressive acts, subject to stable and widely agreed upon norms.

* University of Richmond.
** Hebrew University of Jerusalem.
We are grateful to participants in the criminal law in online spaces workshop for thoughtful and extensive comments on an earlier draft.

TABLE OF CONTENTS

I. INTRODUCTION

The internet facilitates a great deal of very harmful behavior. Some of that behavior, such as fraud, identity theft, or posting embarrassing images of others, is also morally wrong. Moreover, the internet amplifies bad behavior that already existed: with the rise of social media, the humiliating message scrawled on the bathroom stall has gone global. Furthermore, the internet is not simply a conduit for behavior that would otherwise have occurred in some other venue. It provides opportunities for new types of misconduct, such as trolling strangers across the world or hacking into databases and stealing private information.

With the mainstreaming of online life, as well as the centralization of traffic to a few main sites or apps—Google, YouTube, Facebook, Amazon, Twitter—the heady techno-anarchic days of the early internet have given way to an arguably more mature, but also more risk-averse cultural moment. Social media, in particular, has come in for heavy criticism, as it has been blamed for contributing to a wide range of social ills, from the rise in political polarization to the mental health crisis among young people to the dissemination of "fake news" about vaccines and public health to threatening liberal democracy itself.[1] Social media is also blamed for more retail harms, such as generating high volumes of misogynistic, racist, or otherwise hateful invective, radicalizing aimless children, and encouraging users to "pile on" with increasingly heated and vitriolic rants, all in the name of user "engagement."[2]

1. Jonathan Haidt has been particularly vocal in drawing a causal link between social media and poor mental health, particularly among teenage girls. For a brief overview of Haidt's position, see Jonathan Haidt & Nick Allen, *Scrutinizing the Effects of Digital Technology on Mental Health*, 578 NATURE PORTFOLIO 226, 226–28 (2020); for a recent statistical analysis, see Jean M. Twenge, Jonathan Haidt, Jimmy Lozano & Kevin M. Cummins, *Specification Curve Analysis Shows That Social Media Use Is Linked to Poor Mental Health, Especially Among Girls*, 224 ACTA PSYCHOLOGICA 1 (2022).

2. Joshua A. Tucker et al., *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature*, HEWLETT FOUND. (Mar. 21, 2018), https://www.hewlett.org/wp-content/uploads/2018/03/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf [https://perma.cc/YB25-46QH]; Sang Ah Kim, *Social Media Algorithms: Why You See What You See*, 2 GEO. L. TECH. REV. 147, 148–49 (2017); Issie Lapowsky, *Eric Schmidt: Social Media Companies 'Maximize Outrage'*

Unsurprisingly, tech companies, and again social media companies especially so, have come under significant pressure to more aggressively "moderate" their users' speech, for instance, by flagging statements they deem to be misleading or false, suppressing recommendations for inflammatory speech on their algorithms, and temporarily or permanently banning users.[3] Also, unsurprisingly, these calls have generated pushback on free speech grounds, particularly in high-profile and high-stakes political controversies, such as the decision to kick Donald Trump off of Twitter or to suppress stories pertaining to Hunter Biden's laptop during the 2020 presidential campaign.[4] In response, critics argue that social media sites have evolved into something akin to an online "public forum," implying that constitutional limits on the government's ability to restrict speech should be applied to private actors as well.[5]

The regulation of online speech gives rise to questions that have been discussed extensively in the past. Yet, as we show in this paper, it also gives rise to new challenges that differ fundamentally from the questions discussed in traditional free speech jurisprudence. In this paper, we focus attention on the apparently more technocratic question of how user speech should be regulated. We distinguish between an *ex ante* "prevention" model and an *ex post* "punishment" model and consider the conditions that favor the former and those that favor the latter. Our approach is only apparently technocratic, however, as the choice of means is itself heavily value-laden. Our aim is to unpack and partially defend some of those values. Nonetheless, a feature of our approach is that it does not rest on strong presuppositions about the nature and stringency of free speech rights. Our approach should be of interest both to those who consider themselves free speech "absolutists" as well as to those who believe free speech rights must be balanced against a wide range of other rights and interests.

The first part of the paper is devoted to identifying four principal factors relevant to the choice between *ex ante* and *ex post* modes of regulation: fit, error costs, transparency, and normative adaptation.[6] While the distinction between *ex ante* prevention and *ex post* punishment is broadly applicable throughout the law, the second part of the paper is devoted to explaining why the value of normative adaptation is of special significance when it comes to regulating online speech.[7]

for Revenue, PROTOCOL (Jan. 6, 2022), https://www.protocol.com/bulletins/eric-schmidt-youtube-criticism [https://perma.cc/L8Q8-2H3U].

3. For various moderation definitions, techniques, and case studies, see James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42 (2015).

4. *See generally Permanent Suspension of @RealDonaldTrump*, TWITTER (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension [https://perma.cc/59TG-2QJE]; Farnoush Amiri & The Associated Press, *Ex-Twitter Executives to Tell Congress Why They Blocked the Hunter Biden Story in the Weeks Before the 2020 Election*, FORTUNE (Jan. 30, 2023, 1:12 PM), https://fortune.com/2023/01/30/twitter-hunter-biden-story-executives-congress-testify-2020-election-joe-biden/ [https://perma.cc/VMC4-2LPD]; Aja Romano, *Kicking People Off Social Media Isn't About Free Speech*, VOX (Jan. 21, 2021, 3:30 PM), https://www.vox.com/culture/22230847/deplatforming-free-speech-controversy-trump [https://perma.cc/26ZL-EEZT].

5. *See* Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353, 1358 (2018).

6. *See infra* Part II.

7. *See infra* Section III.D.

Although welfarist considerations of fit and error cost tend to support more aggressive *ex ante* prevention, we argue that the importance of normative adaptation weighs against wholesale prevention of undesirable online speech.[8] Normative adaptation suggests that there can be positive value associated with norm-violating speech. More particularly, violations of norms facilitate (and even force) the periodic re-evaluation of the desirability of these norms. The need to punish violators *ex post* provides an opportunity to evaluate our continued commitment to the norm in question.

II. EX ANTE, EX POST

Suppose A is about to assault B. Suppose, further, that you could press a button that would magically prevent A from doing this. Pushing the button would not cause A pain, horrific mind control, or anything like that. Pushing the button just selects the possible world in which A loses interest in assaulting B. Why *wouldn't* you push the button? If you don't push the button, not only will B suffer the assault, but A is also likely to be arrested, prosecuted, and punished for assaulting B. So, both A and B are better off if you push the button. Pushing the button protects B's rights, and it is not obvious that it violates any of A's rights. True, by pushing the button, you limit A's autonomy by taking away A's option to choose not to assault B. On the other hand, since assault is a crime, A doesn't have a legitimate interest, much less a right, to choose to assault B.[9]

So far, so good. But the same reasoning suggests you ought to press the button when doing so would magically prevent *all* assaults. After all, why would we choose a world with both more victims and more punishment when we could choose a world that had dramatically less of both? Again, assailants would not be harmed in any way. And the button is infallible! If it predicts assault, then someone is really in for it, and if someone is really in for it, then the button predicts it. And yet, or so we assume, it does not seem unreasonable to be less enthusiastic about button-pushing in this case than in the previous one.

Here is one reason you might hesitate: norms, including legal norms, are often overbroad. For instance, in Canadian law, assault is defined as intentional but unconsented touching.[10] This is obviously a very broad category; indeed, it is wildly over-inclusive, encompassing slaps on the back, tapping someone's shoulder, unsolicited hugs, bumps during a pick-up game of basketball, and a myriad of other forms of everyday conduct that do not ordinarily merit a criminal conviction, much less punishment.[11] The statutory prohibition also does not take into consideration contextual factors, such as the relationship of the parties,

8. *See id.*

9. There are theorists who argue that, at times, we have 'a right to do wrong.' *See* Jeremy Waldron, *A Right to Do Wrong*, 92 ETHICS 21, 39 (1981). Waldron focuses his attention on moral rights rather than, as in our example, legal rights.

10. Canada Criminal Code, R.S.C. 1985, c C-46, s.265(1)(a).

11. *See generally id*; Piotr Bystranowski & Murat C. Mungan, *Proxy Crimes*, 59 AM. CRIM. L. REV. 1, 12 (2021).

whether the touching occurred during a sports match, or local customs (*e.g.*, about slapping co-workers on the back), which inform most commonsense judgments as to whether an intentional and unconsented to touching merits the label of an "assault."[12] Consequently, if pushing the button prevents all "assaults," thus defined, not only will it stop A from punching B, but it will also stop A from giving B a comforting hug, a pat on the back, or tousling his hair. Maybe on balance, you should still push the button, but that decision now calls for a careful weighing of claims and interests, whereas the decision to push the button in the first case does not.

Our aim in this paper is, first, to provide a typology of reasons why you might hesitate about pushing the button in the second case even if you would not hesitate in the first case, and second, to explore how that typology applies in the context of content moderation by social media platforms. Pushing the button stands for the idea of *ex ante* regulation, which seeks to avoid some identified harm by preventing people from causing it in the first place.[13] For instance, a censor may block a website from posting an embarrassing story or risqué videos. Since *ex ante* regulation seeks to remove an option from an agent's set of possible actions, we refer to it as 'prevention.' Not pushing the button, but rather allowing people to act and then punishing them if they chose poorly, stands for the idea of *ex post* regulation.[14] Because criminal law is the paradigmatic instance of *ex post* regulation, we regard 'punishment' as the paradigmatic form of *ex post* regulation. As we establish later, the distinction is not entirely neat and clear-cut, and there may be intermediate cases.

Online content moderation can operate either *ex ante* or *ex post*. Automated detection of obscenity, nudity, violence, and other prohibited categories of speech may be imperfect, but nonetheless, their aim is to deprive people of the very power to use social media as a platform for publicizing prohibited categories of speech.[15] In contrast, Facebook's "Supreme Court" or the European "right to be forgotten" operates on an *ex post* basis, in that a decision whether to delist or remove speech occurs only after that speech has already been made.[16] Sometimes it can straddle the line, for instance, by allowing users to flag suspect speech leading to its expeditious removal: formally *ex post*, but approximating *ex ante*,

12. *See generally* Canada Criminal Code, R.S.C. 1985, c C-46, s.265(1)(a); Bystranowski & Mungan, *supra* note 11, at 12.

13. *See Ex Ante*, BLACK'S LAW DICTIONARY (11th ed. 2019); Charles D. Kolstad, Thomas S. Ulen & Gary V. Johnson, *Ex Post Liability for Harm vs. Ex Ante Safety Regulation: Substitutes or Complements?*, 80 AM. ECON. REV. 888, 888 (1990).

14. *See Ex Post*, BLACK'S LAW DICTIONARY (11th ed. 2019); Kolstad, Ulen & Johnson, *supra* note 13, at 888.

15. Online content moderation has thus been characterized as both "a vast system of prior restraint" and "indispensable to Internet communications." *See* Langvardt, *supra* note 5, at 1357.

16. *See* HERKE KRANENBORG, *Article 17. Right to Erasure ('Right to Be Forgotten')*, *in* THE EU GENERAL DATA PROTECTION REGULATION (GDPR): A COMMENTARY 475–84 (Christopher Kuner, Lee A. Bygrave & Christopher Docksey eds., 2020); Kate Klonick, *Inside the Making of Facebook's Supreme Court*, NEW YORKER (Feb. 12, 2021), www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court [https://perma.cc/ARS2-CHAA].

depending on how quickly speech is flagged and removed. Moreover, *ex ante* and *ex post* regulation can operate iteratively. For instance, common law judicial systems that respect the norm of *stare decisis* initially resolve controversies *ex post*, but *stare decisis* means that those decisions have an *ex ante* effect in the future by preventing the re-litigation of the same issue.

The case for *ex ante* regulation is intuitive: it prevents the circulation of incendiary, false, misleading, humiliating, or otherwise bad speech, thus mitigating the wrongs and harms said to flow from such speech. This is, one might well think, just like pushing the button to prevent A from punching B. But when *ex ante* content moderation works by blocking entire categories of speech acts, it becomes analogous to pushing a button to block *all* assaults rather than a specific one. And just as one might wonder whether the case for *ex ante* prevention is weaker when it comes to preventing all instances of assault as opposed to a specific one, one might have similar doubts when it comes to *ex ante* content moderation.

III. FIT, ERROR COST, TRANSPARENCY, AND NORMATIVE ADAPTATION

We do not seek to provide a comprehensive theory of the relative merits of *ex ante* and *ex post* regulation. Instead, we identify four broad factors that should be highly salient in assessing a choice between *ex ante* and *ex post*. Three of the categories—fit, the cost of precautions, and transparency—are reasonably straightforward. The last factor, norm adaptation, is perhaps somewhat less so.

A. Fit

Fit refers to the degree to which a norm avoids both over- and under-breadth, that is, the degree to which it includes all and only the actions we wish to regulate.[17] Fit is thus a measure of how well we can distinguish between actions (or omissions) that ought to be prohibited and actions that ought not to be prohibited in advance.[18]

Often, in order to address concerns of over- or under-breadth, the law uses standards rather than rules. For instance, Canada's Criminal Code prohibits any public communication that "wilfully promotes hatred against any identifiable group."[19] What counts as hatred, and what, specifically, does a statement have to be like to "promote" it? We can surely all think of examples of speech we would consider to be "hate speech," but explaining what those examples have in common is difficult, if not impossible. So perhaps it is unsurprising that the statutory language just isn't that precise.[20]

17. *See* JOHN STUART MILL, ON LIBERTY 72 (1859).

18. The argument from fit goes back at least to Mill. *See id.* at 88.

19. Canada Criminal Code, R.S.C. 1985, c C-46, s.319(2).

20. For a discussion of the problems in identifying what hate speech is, see Alon Harel, *Hate Speech*, *in* THE OXFORD HANDBOOK OF FREEDOM OF SPEECH 455–76 (Adrienne Stone & Frederick Schauer, eds., 2021).

Sometimes, there are good reasons for using standards. The meaning of an action may be highly context-specific, as in the case of an embrace. Similarly, whether speech is hate speech may turn on context and circumstances that are difficult or impossible to specify clearly in a rule. The U.S. Supreme Court first defined fighting words in *Chaplinsky v New Hampshire* as words which, "by their very utterance, inflict injury or tend to incite an immediate breach of the peace."[21] But whether the utterance of a string of words rises to that level is highly context-specific, as our starting examples illustrate. Of course, context specificity comes in degrees. Perhaps in some cases, the salient contexts are characterized by stably recurring patterns, which might be readily specified in a rule. In other cases, however, the range of relevant contextual cues may be vast or indeterminate. In such cases, whether an utterance "inflict[s] injury or tend[s] to incite an immediate breach of the peace" could only realistically be known *ex post*, namely by whether it, in fact, caused injury or incited a breach of the peace.[22]

In other cases, while it might not be *impossible* to create a rule with a high degree of fit, crafting a rule might be more trouble than it is worth—for instance, if the actions we are interested in are very rare, acquiring the relevant information is costly, or crystallizing it in detailed rules is time-consuming. Here, we draw on Kaplow's argument regarding when it would be desirable to expend the effort to devise more complex, but more precise, rules rather than leave things to fuzzy standards.[23] Kaplow's argument turns on whether it is more efficient to concentrate decision-making efforts *ex ante* or to put them off until concrete situations that demand their resolution arise.[24] Thus, in the context of online content moderation, the ubiquity of undesirable speech, as well as technological advances in filtering unwanted speech—particularly if the task of devising increasingly sophisticated filters itself becomes automated—are both of particular importance. Both would tend to justify greater investment in *ex ante* as opposed to *ex post* regulation. Indeed, as we discuss below, the sheer volume of online speech—for instance, over 500 hours of video is uploaded to YouTube every minute[25]—weighs heavily in favor of *ex ante* regulation. This justifies an awful lot of expenditure in crafting highly detailed *ex ante* rules.

Moreover, technological innovation can potentially reduce the cost of distinguishing between permissible and impermissible speech *ex ante*, even if that is difficult to do today. After all, filters that block pornographic websites are

21. *See* Chaplinsky v. New Hampshire, 315 U.S. 568, 572 (1942).

22. This is related to the debate as to whether the test in *Chaplinsky* ought or ought not to take into consideration the circumstances in which the speech was made. *See* Michael J. Mannheimer, *The Fighting Words Doctrine*, 93 COLUM. L. REV. 1527, 1527 (1993).

23. *See* Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L.J. 557, 586–96 (1992).

24. *See id.*

25. *YouTube for Press*, YOUTUBE, https://blog.youtube/press/ (last visited July 4, 2023) [https://perma.cc/2BN3-UMHE]. Similarly astronomical figures have been reported for other social media sites. *See, e.g.*, *The 2014 #YearOnTwitter*, TWITTER (Dec. 10, 2014), https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html [https://perma.cc/AR4J-PCWM].

already ubiquitous.[26] Content-based filtering algorithms could advance to the point that they accurately and precisely distinguish not only pornographic websites but also hateful, threatening, or otherwise impermissible speech.

Summing up, there is a reason to prefer *ex post* to *ex ante* law when, for principled or practical reasons, the best we can manage is drafting a norm with a low degree of fit with the underlying justificatory considerations, whether in the form of an over-broad rule or a vague standard. *Ex post* regulation often provides us the opportunity to distinguish "genuinely" harmful actions from spurious overbreadth. Thus, even if theoretically we could prevent in advance every single technical assault, there is a good reason not to do so since a lot of what is technically an assault is morally innocuous.[27] Hence, in effect, we treat what seems on its face to be a rule as a standard and decide *ex post* whether or not to punish violations. There seems to be little reason to prevent parents from hugging their children, even though such actions are, strictly speaking, intentional and unconsented touchings. Similarly, for irreducibly vague laws, *ex ante* regulation requires the advance specification of actionable rules, precluding case-by-case determination of each potentially impermissible act in the context of its occurrence.[28]

B. Error Cost

Error costs are closely related to fit. *Ex post* regulation incurs obvious costs both in the form of harmful actions that would have been prevented under an *ex ante* rule as well as in the form of *ex post* adjudication. Therefore, if it is desirable to have a tightly fitted rule, then it is probably also desirable to censor rather than punish. Prevention's benefits rise, and its error costs fall, as the norm governing speech becomes more precisely targeted. For instance, if we can reliably distinguish obscenity or disinformation from permitted speech—avoiding both types of error—then there is a case for censoring such speech since we thereby avoid the harms of social media littered with obscenity and disinformation while simultaneously *not* censoring speech that is neither obscene nor disinformation.[29]

26. *See, e.g.*, *Filter Explicit Results Using SafeSearch*, GOOGLE, https://support.google.com/websearch/answer/510 (last visited Feb. 18, 2023) [https://perma.cc/6N3G-9T6E]; Note, *The Impermeable Life: Unsolicited Communications in the Marketplace of Ideas*, 118 HARV. L. REV. 1314, 1332 (2005); Michael L. Rich, *Should We Make Crime Impossible?*, 36 HARV. J.L. & PUB. POL'Y 795, 798 (2012).

27. *See* Rich, *supra* note 26, at 812.

28. *See* Langvardt, *supra* note 5, at 1362; Rich, *supra* note 26, at 812–14. Note that our argument here is conditional on the existence of principled or practical reasons for tolerating a loosely fitted law. Some laws may be more loosely fitted than they should be. Perhaps we are too quick to assume, for instance, that assault ought not to be defined a bit more precisely than it is. We take no stand on those questions here.

29. It is also true that the volume of online speech makes effective punishment *ex post* an impossibility. *Cf. YouTube for Press*, *supra* note 25. But *ex post* regulation need not take the form of criminal punishment. *E.g.*, *YouTube Community Guidelines Enforcement*, GOOGLE, https://transparencyreport.google.com/youtube-policy/removals?hl=en (last visited July 4, 2023) [https://perma.cc/VK3J-SMXU]. Taking down user speech based on complaints is also a form of *ex post* regulation. *E.g.*, *id.*; Grimmelmann, *supra* note 3, at 90.

One might choose to weigh the different types of error (false positives and false negatives) differently, as we do in criminal trials. Doing so might influence the choice between *ex ante* and *ex post* regulation. For instance, consider a series of social media posts encouraging someone to commit suicide. The costs of allowing such speech are potentially extremely high, whereas the costs of erring the other way amount to preventing people from engaging in permissible but low-value speech. Conversely, the case for *ex post* punishment is strengthened if the cost of wrongly censoring innocent speech is substantially greater than that of wrongly permitting impermissible speech. For instance, suppose governments are prone to label the speech of their political and cultural adversaries as "fake news" or "disinformation." In that case, the costs of allowing actual disinformation to percolate on social media are very likely to be dwarfed by the costs of allowing the government to censor speech it finds embarrassing or inconvenient.[30]

The sheer volume of conduct to be regulated—recall that the scale of user speech on social media platforms is staggering[31]—is relevant once again. Even if false positives are typically much worse than false negatives—perhaps people resent being kicked off a platform for no good reason much more than they resent run-of-the-mill trolling—still, if trolling becomes endemic on a platform, that can eventually outweigh the resentment of those few who are inadvertently blocked. In the case of online content moderation, an algorithmic filter that occasionally blocks innocent speech may be superior, on welfarist grounds, to an *ex post* system that cannot keep pace with the volume of illicit speech and thus winds up effectively leaving people wronged by such speech without practical recourse—even if mistakenly excluding someone from an online platform is, in isolation, a more serious wrong.[32]

Further, digital platforms often have very large audiences, which may weigh in favor of heavier reliance on *ex ante* prevention.[33] The ease and immediacy of publication that is often done impulsively; the fact that publication is spread quickly and, further, the broad geographic and temporal reach of online speech may imply that the damage resulting from speech can be quite substantial.[34] All these considerations suggest that *ex ante* regulation designed to prevent the speech from ever being made may be preferable.[35]

30. *See, e.g.*, Anton Troianovski & Valeriya Safronova, *Russia Takes Censorship to New Extremes, Stifling War Coverage*, N.Y. TIMES (May 18, 2022), https://www.nytimes.com/2022/03/04/world/europe/russia-censorship-media-crackdown.html [https://perma.cc/U2GX-PREK]; Gertz v. Robert Welch, 418 U.S. 323, 340–41 (1974).

31. *See, e.g.*, *YouTube For Press*, *supra* note 25.

32. *See* Langvardt, *supra* note 5, at 1357, 1360–62; Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1626–27 (2018).

33. *See* Ariel L. Bendor & Michal Tamir, *Prior Restraint in the Digital Age*, 27 WM. & MARY BILL RTS. J. 1155, 1172 (2019).

34. *Id.* at 1170–71.

35. *See id.* at 1170–74.

Asymmetric precautionary costs provide one way to rationalize the American view of defamation. Public figures may suffer morally impermissible harm to their reputation from the activities of muckraking journalists and internet pundits, but (one might argue) it would be worse still to chill commentators from speaking freely on matters of public importance.[36] If that view is correct, it would make sense to allow people to speak freely about public figures while giving such figures an *ex post* remedy of damages in those cases where the speaker acted with malice.[37] On the other hand, if toleration of false, reputation-damaging speech is a greater evil than chilling open public discourse, then *ex ante* prevention is the preferred regulatory option. Better to have Facebook, Twitter, et al. preemptively block potentially defamatory statements at the outset rather than require people to sue after their reputations have been tarnished.

The regulation of pornographic images of children provides a second example. The sharing of such images constitutes a grave wrong against their subjects, one that is not undone by the punishment of those responsible.[38] There are substantial costs to blocking people from sharing images that are erroneously flagged as child pornography. In a famous case, Facebook censored the picture of Kim Phuc fleeing from a Napalm attack during the Vietnam War.[39] Plausibly, this was a bad call: this photo is not pornographic but rather serves the purpose of demonstrating the horrors of war. One can imagine less dramatic cases as well, such as a grandparent who is frustrated, and more than a little offended, by being prevented from sharing photos of a newborn grandchild because an algorithm flags the photo as child pornography. How one weighs these different types of error requires a substantive moral judgment. Some would argue that it may be better to wrongly prevent a few innocent photos rather than rely on a system where the only remedy for exploited children is criminal prosecution.[40] Others may believe that preventing grandparents from posting naked pictures of their grandchildren is a very costly restriction. And, of course, base rates matter: how

36. *See* Erik Walker, *Defamation Law: Public Figures—Who Are They?*, 45 BAYLOR L. REV. 955, 956–57 (1993).

37. *Id.* at 979–81. If the distinction between private and public figures is highly context-specific, then (per our previous argument) there is further reason to favor a more limited *ex post* regime. *See id.* at 979.

38. *See Child Pornography*, U.S. DEP'T OF JUST., https://www.justice.gov/criminal-ceos/child-pornography (May 28, 2020) [https://perma.cc/H99N-8NAD]; Mary M. Giannini, *Slow Acid Drips and Evidentiary Nightmares: Smoothing Out the Rough Justice of Child Pornography Restitution with a Presumed Damages Theory*, 49 AM. CRIM. L. REV. 1727, 1727–30 (2012).

39. Sam Levin, Julia Carrie Wong, & Luke Harding, *Facebook Backs Down from 'Napalm Girl' Censorship and Reinstates Photo*, GUARDIAN (Sept. 9, 2016, 1:44 PM), https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo [https://perma.cc/JG23-DNPG]; Kjetil M. Hovland & Deepa Seetharaman, *Facebook Backs Down on Censoring 'Napalm Girl' Photo*, WALL ST. J. (Sept. 9, 2016, 3:07 PM), https://www.wsj.com/articles/norway-accuses-facebook-of-censorship-over-deleted-photo-of-napalm-girl-1473428032?page=1 (last visited July 4, 2023) [https://perma.cc/CW2K-Q5XQ].

40. *See generally* Amy Adler, *Inverting the First Amendment*, 149 U. PA. L. REV. 921 (2001) (arguing that child pornography statutes are overbroad and have strayed from their original purpose); *see also* Langvardt, *supra* 5, at 1361–62. This is not to say, of course, that people caught disseminating child pornography ought not be punished but rather that the legal system's main efforts in this area should be directed at trying to prevent the dissemination of such images in the first place. *See* New York v. Ferber, 458 U.S. 747, 760 (1982).

often are people uploading exploitative versus innocent photos? Our point is only that the content of that judgment will affect the case for *ex ante* or *ex post* means of regulating speech.

C. Transparency

Ex ante prevention is often less publicly observable than *ex post* punishment, although this is a rough generalization rather than a categorical rule.[41] Because *ex post* approaches tend to require case-specific judgment, they are probably more closely associated with traditional forms of legal adjudication.[42] In contrast, *ex ante* approaches are more naturally associated with rule-based regulation.[43] Moreover, the vivid detail and evidence generated by concrete actions and realized harms are likely to generate greater public attention than the more abstract discussion of general rules and practices.[44] For instance, the phenomenon of unjustified police use of force was widely known and discussed among specialists, but it took a video of a murder committed by American police to effectively galvanize public sentiment.[45]

While there are many exceptions—administrative rulemaking in the United States can sometimes generate significant volumes of public comment—in the context of online content moderation, *ex ante* regulation is mostly a matter of internal company policy, which can be very hard for outsiders to observe in any reasonably granular manner.[46] This can be valuable insofar as intense public scrutiny and controversy impair effective regulation or are simply socially wasteful.[47] Hence, in cases where it is clear that an *ex ante* strategy is preferable, the relative lack of scrutiny of particular claims and factual contexts is probably a feature rather than a bug. It is probably for the best, for instance, to adopt a largely *ex ante* approach to child pornography since there is little social value to publicly debating the precise contours of that norm, meaning that the norm can be formulated vaguely ('we know it when we see it'), leaving it to platforms to develop more granular *ex ante* policies behind the scenes. Or, to take an example relating to street crime (as opposed to speech), assume that the evidence that

41. *See* Rich, *supra* note 26, at 828; Grimmelmann, *supra* note 3, at 65.

42. *See* Klonick, *supra* note 32, at 1622; Kaplow, *supra* note 23, at 582–85; Grimmelmann, *supra* note 3, at 67.

43. *See* Klonick, *supra* note 32, at 1631–32; Kaplow, *supra* note 23, at 559–60.

44. *See* Langvardt, *supra* note 5, at 1361–62; Lani Guinier & Gerald Torres, *Changing the Wind: Notes Toward a Demosprudence of Law and Social Movements*, 123 YALE L.J. 2740, 2758 (2014).

45. The value of transparency is also likely correlated with fit: substantial moral disagreement can sometimes be the cause of vague or overbroad laws, with more granular—and hence, controversial—decisions postponed from the norm-making to the norm-applying stage. *See* Derrick Bryson Taylor, *George Floyd Protests: A Timeline*, N.Y. TIMES (Nov. 5, 2021), https://www.nytimes.com/article/george-floyd-protests-timeline.html [https://perma.cc/33RU-49U3].

46. Langvardt, *supra* note 5, at 1362; Grimmelmann, *supra* note 3, at 65.

47. *See* Klonick, *supra* note 32, at 1650–58.

street lighting can reduce crime is sound.[48] In that case, streetlights would be an *ex ante* mode of crime prevention, in contrast to the *ex post* criminal law model of responding to crime after the fact. Given our assumption, the planning committee's meeting where the budget for streetlights is determined may result in a better approach to reducing crime than a series of public trials for people caught committing crimes at night. That the trials are more public and transparent than a routine planning committee meeting does not mean that they will result in better policy.

But this reasoning can be pernicious when the substantive values underlying a preference for *ex ante* prevention are themselves the subject of reasonable disagreement.[49] Transparency takes on greater significance in such cases for standard democratic reasons. As Mill put it, while it is "undisputed" that government should make efforts to prevent crime,

> [t]he preventive function of government … is far more liable to be abused, to the prejudice of liberty, than the punitory function;—for there is hardly any part of the legitimate freedom of action of a human being which would not admit of being represented, and fairly too, as increasing the facilities for some form or other of delinquency.[50]

Given that it is quite easy to create a facially plausible "harm prevention" rationale for an indefinitely large range of acts, insisting on a public hearing and punishment *ex post* has positive value in providing a venue for checking regulatory power. That said, *ex ante* regulation may be preferable in contexts where ground truth is readily established; it is less troubling to censor people who insist the earth is flat or the moon is made of green cheese than it is to censor people who question prevailing norms about, say, equality.

Moreover, transparency can be significant even in cases where the underlying value is not itself controversial. The controversies engendered by *ex post* adjudication and punishment can serve a norm-revitalizing function, as they arguably did in the case of the recent Black Lives Matter rallies.[51] By providing an occasion for collective condemnation of norm-violative acts, *ex post* punishment is a means for strengthening social norms and collective solidarity. Consider hate speech again. Even if we could cleanly distinguish genuine hate speech from merely offensive or boundary-pushing speech, we might not wish to prevent hate speech entirely. At least, not if the prosecution, trial, and punishment for someone who directs hatred at another has public educative value, for instance, by reminding us of why we have a norm against hate speech or what the content of

48. Aaron Chalfin, Benjamin Hansen, Jason Lerner & Lucie Parker, *Reducing Crime Through Environmental Design: Evidence from a Randomized Experiment of Street Lighting in New York City*, 38 J. QUANTITATIVE CRIMINOLOGY 127, 151 (2012).

49. *See* Bendor & Tamir, *supra* note 33, at 1167–68.

50. JOHN STUART MILL, ON LIBERTY 88 (1859). *See also* Rich, *supra* note 26, at 828.

51. *See* Taylor, *supra* note 45; Ram Subramanian & Leily Arzy, *State Policing Reforms Since George Floyd's Murder*, BRENNAN CTR. (May 21, 2021), https://www.brennancenter.org/our-work/research-reports/state-policing-reforms-george-floyds-murder [https://perma.cc/WG3S-8NBN].

that norm is (or should be.)[52] Without some quantum of *ex post* adjudication, the perniciousness of hatred may be forgotten and the pains resulting from hate speech ignored or underappreciated. Punishing those who engage in hate—not merely censoring them—thus can serve as a type of solidarity-building or educative exercise.[53]

The norm-reinforcing theory of punishment is most famously associated with Durkheim, who observed that it predicts a kind of penal homeostasis, whereby groups have reason to keep levels of symbolically charged punishment constant, even in the face of declining rates of norm violation.[54] In Durkheim's view, this creates a steady demand to find transgressions of shared norms so that we can demonstrate in-group solidarity through public condemnation and punishment.[55] Obviously, the value of transparency in revitalizing flagging norms is limited. At some level, it begins to undermine liberal values, as it suggests sacrificing individuals in the name of building collective sentiment. Moreover, the moral value of transparency is entirely dependent upon the underlying moral value that one seeks to revitalize. Despots know that starting a war can rally public sentiment, just as a fading celebrity knows that causing an outrageous public scene will help her stay in the public eye. That does not redound to their moral credit.

D. Normative Adaptation

Normative adaptation is a gauge of a normative system's ability to adapt to changing circumstances and values.[56] Normative adaptation has two dimensions: adaptation's *value* and adaptation's *agent*. On the first, whether adaptation or rigidity is preferable is a function of how wedded we are to a particular norm. That can be intuitively assessed by asking how difficult it would be to imagine a context in which we would abandon the norm and adopt a fundamentally different one. For instance, it is quite hard to imagine circumstances in which our judgment about the permissibility of child pornography would flip. In that case, we might not value norm adaptation very much; in fact, we might positively disvalue it as a form of moral worsening.

In other cases, however, our commitment to a norm is either more ambiguous or very abstract, concealing a lot of disagreement and uncertainty in particular cases. In those cases, there is reason to prefer a less rigid system of norms that can adapt as we refine our attitudes and knowledge. Conflicts can, as Nils

52. *See* Langvardt, *supra* note 5, at 1385. Thus, per Beccaria, "the severity of a punishment should be just sufficient to excite compassion in the spectators, as it is intended more for them than for the criminal." CESAR BONESANA, AN ESSAY ON CRIMES AND PUNISHMENTS 100 (Edward D. Ingram trans., 1819) (1762).

53. Rich, *supra* note 26, at 825.

54. EMILE DURKHEIM, THE RULES OF SOCIOLOGICAL METHOD 115–16 (1st Am. ed., Steven Lukes, ed., W. D. Halls, trans., Free Press 1982) (1895).

55. *Id.*

56. Thomas W. Platt, *Adaptation as a Normative Concept*, 80 ETHICS 230–34 (1970).

Christie famously argued, be opportunities for norm clarification.[57] For example, in many of the types of speech that are at the center of current controversies over content moderation online—such as inflammatory political speech, disinformation, and offensive speech in a wide variety of guises—the line between the merely controversial and the truly intolerable is itself very uncertain.[58] This counsels against rigidifying current attitudes, no matter how intensely held. After all, many now accepted forms of scientific, artistic, political, and cultural expression were at one point highly controversial and reviled.[59]

Assuming that there is a positive value to norm adaptation, the second question is the source of normative adaptation, or in other words, adaptation's agent. *Ex ante* regulation can be changed but tends to be top-down (bureaucratic), whereas *ex post* regulation tends to be bottom-up (judicial).[60] Suppose a social media platform adopts a rule preventing people from expressing certain thoughts or sentiments on their platform. Changing that rule requires appealing to the rule-makers at the company. Absent some other less regulated platform, people whose (potentially inflammatory) posts never see the light of day will have a harder time drawing attention to themselves than they would if the speech was first published and then removed or sanctioned after some kind of case-specific decision. This suggests that in many cases, *ex ante* regulation will likely prove to be more rigid than *ex post* punishment, on the assumption that *ex post* punishment enjoys a somewhat higher degree of visibility. That said, *ex ante* systems may also be more brittle in the sense that if there is a change in policy or personnel, *ex ante* rules are liable to more rapidly change than the more incremental, evolving character of case-by-case judgments *ex post*.

Without seeking to propose a systematic theory of norm change, we offer the following tentative observations. First, top-down adaptation is a virtue in contexts in which there is good reason to trust expert opinion when it diverges from lay opinion, as the rigidity insulates unpopular, but substantively correct, decisions from less informed lay judgment. Insofar as experts are well-chosen and acting within the scope of their expertise, this should be true in most cases. Top-down adaptation is a virtue in information-rich contexts where the rule-makers update their views regularly, as it can, in principle, lead to much more rapid adaptation than decentralized case-by-case judgment. On the other hand, bottom-up approaches enjoy an advantage when there is a high risk of bureaucratic stasis, as responding to cases as they arise provides opportunities for jump-starting a sclerotic bureaucracy.[61] Bottom-up approaches are also advantageous in contexts

57. Nils Christie, *Conflicts as Property*, 17 BRIT. J. CRIMINOLOGY 1, 8 (1977).

58. *See* Lauren Feiner, *How the Supreme Court Could Soon Change Free Speech on the Internet*, CNBC (Jan. 30, 2023, 12:06 PM), https://www.cnbc.com/2023/01/30/the-supreme-court-could-change-free-speech-on-the-internet.html [https://perma.cc/7TN2-5U7N].

59. *See generally* Christie, *supra* note 57.

60. *See* discussion *supra* Part II.

61. *Ex post* punishment is a means of exercising "voice" in forcing an institution to become more responsive. "Exit" options are also possible, in extreme circumstances, by simply bypassing the frozen institution. *See* MARIANA MOTA PRADO & MICHAEL J. TREBILCOCK, INSTITUTIONAL BYPASSES 48 (2018).

where the danger of intellectual fads among elites is high. In those contexts, more decentralized systems may be preferred precisely because they tend to be slower to succumb, at least if the decentralized decision-makers are insulated from outside pressures. A more open forum for challenging received wisdom is valuable in avoiding 'the emperor has no clothes' situations, in which peoples' judgments are more responsive to their beliefs about other peoples' judgments than to their sense of what is actually true.[62]

Second, when it comes to adaptation's value, fallibilism about moral judgment means that even our strongest moral attachments may potentially come in for revision. Obviously, people are sometimes attached to norms that, upon reflection, do not merit such attachment. Only slightly less obviously, occasionally, those people are us. Thus, while the value of normative adaptation may asymptotically approach zero, fallibilism suggests that it should not be discounted entirely. *Ex post* punishment provides a means for continually testing, and refining, *ex ante* rules in the crucible of novel and unexpected contexts.

To be clear, dissenting views are often wrong, so tolerating them means tolerating a landscape that is epistemically more fraught than one pruned of falsehoods and baleful heterodoxies. But dissenting views are also a source of variation in social attitudes that can have salutary effects, from challenging a stifling status quo to deepening appreciation for mainstream norms. Indeed, part of the reason that dissenting views should be expected to be wrong is that self-correcting social attitudes are likely to have already incorporated whatever information they convey. Our Millian point is that social attitudes can only be self-correcting if they are provided with the opportunity and flexibility to update when presented with new information or perspectives that challenge the status quo.

Stepping back from the factors that influence the importance of normative adaptation in a particular context, there is a somewhat larger point in the offing here as well. This is that normative adaptation identifies a distinctive and *positive* value to norm-violative conduct. That is, not only do people sometimes have a right to do wrong, but it is also sometimes *good* for people to act wrongly, even if not for the people they wrong. This is less paradoxical than it seems; indeed, the point is simple: perfect prevention impairs learning and robust discourse.

The typical argument for a right to do wrong is grounded in the importance of autonomy, privacy, or other self-regarding values. In contrast, normative adaptation is a *social* value—it is better for everyone if our norms adapt to changing circumstances rather than remain fixed in a potentially obsolete equilibrium. Normative adaptation thus assigns some positive value to norm-violating conduct. Moreover, because that value is not premised on the existence of the right to do wrong, even if there were no right to do wrong, it would still be good for people to (occasionally) do wrong.

Normative adaptation explains why perfect prevention of wrongful acts is not an unalloyed good. Committing wrongs, especially in the context of speech,

62. CRISTINA BICCHERI, THE GRAMMAR OF SOCIETY 176–213 (2006).

may often contribute to testing the boundaries of what is right. The boundaries between right and wrong in the context of speech are sometimes quite vague and, in any case, are frequently shifting. Testing the boundaries has the potential to contribute to the evolution of moral norms. It improves our qualities as citizens who participate in political discourse. An occasional failure may justifiably expose the violator to moral criticism, but it is not an unmitigated evil. It also contributes to the future drawing and redrawing of the boundaries separating the permissible from the impermissible and the naughty, mischievous, or provocative speech act from the incitement to violence. From Rosa Parks' refusal to give up her seat to the Lovings' determination to marry in the face of Virginia's anti-miscegenation statute to the refusal of gays and lesbians to stay in the closet in the face of state sodomy laws, the civil rights movement is littered with famous instances of disobedience that triggered critical discussion of ossified and oppressive social norms, culminating with their revision.[63] Even apparently obsolete speech norms pertaining to respect for political and cultural figures can be less obsolete than they seem, as demonstrated by the arrest of anti-monarchy protesters following the death of Queen Elizabeth.[64] In such cases, the educative and adaptive rationales for *ex post* regulation are particularly compelling, as it would be extraordinarily illiberal to prevent anti-monarchist and republican protesters from speaking at all. Of course, destabilizing norms can also be much more equivocal, as, for instance, when Donald Trump publicly called for his political rival to be jailed or when he later denied the legitimacy of the 2020 election.[65]

Finally, as Rich has pointed out, normative adaptation is often indirectly served by norm-violative conduct.[66] This occurs when people engage in indirect civil disobedience, violating one norm in order to convey a message about some other norm or practice.[67] Whether it is burning draft cards, trespassing, blocking traffic, or throwing food in art museums, actively flouting a norm is a time-honored mode of civil disobedience. The value of civil disobedience—which, whether it is over- or underrated, is not zero—is undermined by a system of perfect *ex ante* prevention. While the *ex ante*/*ex post* distinction applies on both sides of the speech/conduct divide, in the case of speech, more porous forms of *ex post* regulation can contribute to the assessment and reassessment of norms.

Let us take stock. We have not sought to provide a comprehensive theory of *ex ante* versus *ex post* modes of regulation but only to identify what we take to be the most salient factors that bear on this dimension of regulatory design.

63. Rich, *supra* note 26, at 827.

64. Annabelle Timsit, *Police Arrest Anti-Monarchy Protesters at Royal Events in England, Scotland*, WASH. POST (Sept. 13, 2022, 10:37 AM), https://www.washingtonpost.com/world/2022/09/13/queen-elizabeth-death-protests-arrest-police/ [https://perma.cc/Q4XM-M6HS].

65. David Jackson, *Trump: Clinton Should Be in Jail, the Election Is Rigged*, USA TODAY (Oct. 15, 2016, 2:35 PM), https://www.usatoday.com/story/news/politics/elections/2016/2016/10/15/donald-trump-maine-new-hampshire/92143964/ [https://perma.cc/K4UY-YJFM].

66. Rich, *supra* note 26, at 827.

67. *See id.*

Additionally, we have not attempted to assign weights to these factors but rather have focused on contextual features of a regulatory problem that would tend to strengthen or weaken the salience of a given factor.

The typology we have articulated can be represented as follows:

Fit: A narrowly tailored rule is more apt for *ex ante* regulation. *Ex post* punishment is more appropriate in contexts where we are committed to using vague standards. The choice between a rule and a standard is sensitive to how often the norm is to be applied, as the higher up-front costs of devising a narrowly tailored rule can be spread over a long run of cases.

Error cost: If the cost of false positives is unusually high, that counts in favor of *ex post* punishment. If the cost of false negatives is unusually high, that counts in favor of *ex ante* regulation.

Transparency: How important is it to periodically re-affirm a norm through some public process? Does the conduct in question rest on claims that can be falsified by epistemically secure methods? The strength of the case for *ex ante* regulation is positively correlated with the observability of ground truth and inversely correlated with the importance of public engagement.

Norm adaptation: How sure are we that a norm is (or should be) entrenched? The strength of the case for *ex ante* regulation is positively correlated with our certainty that a norm should be rigid. Are experts or lay publics more likely to update appropriately when faced with new information? The strength of the case for *ex ante* regulation is stronger if it is more likely that experts will appropriately incorporate new information than lay publics.

IV. APPLYING THE TYPOLOGY TO CONTENT MODERATION

We turn, finally and briefly, to questions of application. We use the case of content moderation on social media as an illustrative case. The considerations discussed above apply both to traditional and to new media platforms, although traditional approaches to content moderation are now being rethought because digital media platforms are quite different from traditional media along numerous dimensions, such as reach and scale.[68]

First, the extraordinary context-specificity of online speech suggests that *ex ante* rules would inevitably be extremely complex. Nonetheless, the sheer volume of speech posted to social media platforms suggests that it would be worth spending the time to devise those rules, in part because adjudicating each potential dispute *ex post* is not remotely feasible.[69] A censorship algorithm could potentially obviate some of these concerns, as an algorithm could operate at scale to pre-screen comments, images, and videos for narrowly defined categories of prohibited speech.

68. *See* Bendor & Tamir, *supra* note 33, at 1164–65.

69. *See YouTube for Press*, *supra* note 25.

Error costs depend upon what types of online speech we have in mind. It is obviously bad to prevent a proud grandparent from posting a photo of her new grandchild, but that is a relatively minor harm compared to the dissemination of child pornography. In contrast, consider efforts to censor "misinformation." Suppose a Ministry of Truth censors comments on social media in order to strain out what it deems to be "misinformation" pertaining to an upcoming election. Perhaps in doing so, it mitigates some of the epistemic problems associated with rampant misinformation, such as unfounded conspiracy theories, fearmongering based on obvious falsehoods, and so forth. At the same time, however, unless it operates completely in secret—and it is hard to see how liberal democracy is at all consistent with a secret Ministry of Truth—those benefits are mitigated by the fact that people know that they are only being shown what the Ministry of Truth regards as acceptable. An intuitive heuristic for estimating the costs of wrongfully suppressing alleged misinformation is to imagine that the Ministry of Truth (or a social media company's algorithm) is taken over by people who have very different views than you do as to what counts as "misinformation."[70]

The ongoing debate as to the causal impact of social media on mental health is quite relevant here. It is quite difficult to prove causality, and at the moment, the jury seems to still be out.[71] If a clear causal pathway can be demonstrated, then there will be a correspondingly stronger case for imposing stronger *ex ante* restrictions, although perhaps mostly in the form of restricting children's access to social media. Conversely, if social media consumption does not affect psychological well-being, then the case for expanding *ex ante* censorship will be weakened.

The value of transparency is similarly context-dependent. Suppose someone posts a message claiming that climate change is fake. How likely is it that a public hearing presenting the scientific evidence for anthropogenic climate change will change minds? If that is not very likely, then there will be little transparency-related reason to prefer *ex post* to *ex ante* regulation, particularly given that the ground truth on this question is both well-established and, moreover, better established through scientific channels than on social media. On the other hand, suppose someone posts a message stating that the urgency of climate change is vastly overrated. (We are assuming that the message is sufficiently flagrant to warrant investigation *ex post*; if it is not, then it presumably should also not be filtered out *ex ante*.) In this type of case, arguably, a careful and fair

70. Consider, for instance, PayPal's recent decision to impose a $2500 fine on users who spread "misinformation." Notably, PayPal did not define what counts as "misinformation," but did claim the power to classify an action as spreading misinformation at its "sole discretion." *See* Xinyi Wan, *PayPal's "Misinformation" Fine Sparks Backlash*, JOLT DIGEST (Nov. 1, 2022), http://jolt.law.harvard.edu/digest/paypals-misinformation-fine-sparks-backlash [https://perma.cc/M9SA-4WXA].

71. See generally Tucker et al., *supra* note 2, for evidence indicating a causal relationship. Other studies have not found evidence that social media consumption affects well-being. *See generally* Avinash Collis & Felix Eggers, *Effects of Restricting Social Media Usage on Wellbeing and Performance: A Randomized Control Trial Among Students*, PLOS ONE (Aug. 24, 2022), https://doi.org/10.1371/journal.pone.0272416 [https://perma.cc/D3CK-UFCF].

response on the merits would have more value than preventing that message from being posted in the first place. Yet one ought to recall that even statements that seem false on their face may turn out to be true in the long run.

Finally, while it is hard to imagine a world in which it would be morally permissible to disseminate what we currently regard as child pornography, a great deal of other online speech is evidently tied to contingent and context-specific normative expectations, such as those about when someone's comments cross the line that separates "bad taste" from "harassment," or "abrasive" from "intimidating." This is an area of social life where it is quite easy to imagine significant changes in norms, if for no other reason than that the technology itself is constantly evolving. Privacy limitations, the ability to retroactively wipe images or text from the internet, shifting norms about sex and sexuality—all of these can, and have, affected the morality of online speech. The value of providing occasions for norms about online conduct to adapt and evolve is thus much higher than in the case of street crime.

Relatedly, strong *ex ante* speech regulation amounts, in effect, to giving executives at Twitter, Facebook, and TikTok the power to shape cultural norms about everything from politics to sex to identity. From the point of view of norm adaptation, the question is whether we have reason to trust that those executives (or, for that matter, government bureaucrats in a Ministry of Truth) will make better decisions about what those norms should be than the decisions that would emerge from a decentralized process in which the permissibility of particular instances of speech is considered in context, *e.g.*, through Facebook's so-called "Supreme Court." When what counts as "better" is itself deeply contested, it is desirable to make decisions incrementally and in context rather than in a top-down and categorical manner.

Norm adaptation is thus an important concern that lends support to *ex post* regulation and has a particular relevance to social media platforms, given their role in facilitating both norm transgression and norm enforcement. Moreover, the global reach of social media platforms adds a further layer of concern. Traditionally, a jurisdiction could issue *ex ante* regulation, but citizens would still be able to access transgressive speech in other jurisdictions. What makes the concern of over-rigidity particularly grave in the online context is the fact that social media platforms are global, and hence comprehensive *ex ante* regulation can more effectively block the very possibility of norm adaptation. We suggest that this difference is qualitative rather than quantitative. Insofar as speech restrictions operate globally, they should be evaluated very differently from those that apply only in a particular jurisdiction.

It is not impossible that the social media environment could adapt such that different platforms adopt different content moderation policies and market themselves accordingly. This has been observed with platforms, such as Parler or Truth Social, that target right-wing constituencies. One might thus argue that norm adaptation could be consistent even with quite aggressive levels of censorship via the mechanism of market segmentation, *i.e.*, different groups of people

flocking to different platforms according to their preferences about more free-wheeling versus more sanitized content or for content that is filtered for different types of speech. It would not be surprising if social media platforms adopted a similar strategy. After all, market segmentation is familiar from legacy media outlets, *e.g.*, the New York Times versus Fox News.[72]

That said, the very possibility of market segmentation along these lines presupposes a strong commitment to traditional liberal norms. Social media platforms can compete on the basis of their content moderation policies only if internet service providers, payment processors, financial institutions, and other companies are allowed to sell their services to social media platforms regardless of the content of the speech posted to those platforms. Or, to take another example, market segmentation only allows pornographic and family-friendly websites to co-exist insofar as the expectations of the latter are not imposed in such a way as to make it impossible for the former to continue in business. Thus, the prospect of social media companies competing on the basis of their content moderation policies does not show that norm adaptation can co-exist with aggressive censorship of speech across the board. On the contrary, the very possibility of this type of competition presupposes a robust norm against censorship. Alternatively put, if there is a clear case that we should prevent certain types of speech from being expressed at all, then it is not clear why we should prevent it from being posted to Facebook while allowing it to be posted on 4Chan. It is not as if child pornography is okay if it is posted on a pornographic website, just not on YouTube. To the degree that we are certain that kind of speech is best prevented from existing, then there is little reason to stop at the level of a particular website or platform. The choice between *ex post* and *ex ante* regulation is thus inescapable.

72. Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley & Katerina Eva Matsa, *Political Polarization & Media Habits*, PEW RSCH. CTR. (Oct. 21, 2014), https://www.pewresearch.org/journalism/2014/10/21/political-polarization-media-habits/ [https://perma.cc/SP79-2VPZ].