European Journal of Law and Economics (2020) 50:437–449
https://doi.org/10.1007/s10657-020-09653-7

Optimal social media content moderation and platform immunities

Frank Fagan[1]

Published online: 9 May 2020

Abstract
This article presents a model of the lawmakers' choice between implementing a new content moderation regime that provides for platform liability for user-generated content versus continuing platform immunity for the same. The model demonstrates that lawmakers prefer platform immunity, even if incivility is increasing, if the costs of implementing a platform liability regime are greater than the costs of enforcing status quo law. In addition, inasmuch as implementation of a platform liability regime is coupled with new speech restrictions that are unconstitutional or prohibitively costly, lawmakers prefer immunity, but platforms are free to set strong content moderation policies consistent with existing law.

Keywords Social media · Fake news · First amendment · Section 230 · Communications Decency Act

JEL Classification K16 · K23 · K24 · L82

1 Introduction

Nearly every government, regardless of political model or ideological orientation, is today concerned with fake news and the ostensibly elevated combativeness of media discourse. New concerns with fake news and discursive conflict reflect, in part, the structural changes in news content, delivery, and interpretation enabled by social media (Fagan 2018). Lawmakers are faced with the decision to implement new rules or continue enforcing status quo law. While the status quo in most jurisdictions generally provides for platform immunity for user-generated content, lawmakers can

✉ Frank Fagan
frank.fagan@edhec.edu

[1] EDHEC Business School, Roubaix Cedex 1, France

nonetheless regulate platforms through the enforcement of other rules. For instance, the U.S. Federal Election Campaign Act (FECA) prohibits willful causation of campaign advertisement purchases by non-U.S. persons or entities.[1] Platforms that knowingly induce purchases of campaign advertisements from foreign persons or organizations may be subject to criminal liability. Enforcement of other rules that govern platforms, even if liability is not dependent upon user speech, can moderate platform content inasmuch as compliance with other rules impacts prevailing levels of platform content and civility. Other rules that govern platforms—including competition rules, various privacy laws such as the E.U. General Data Protection Regulation (GDPR), and rules that require expedient notification of data breaches—also moderate content inasmuch as they change the incentives for hosting it.[2] Further, lawmakers can directly prosecute users when they engage in prohibited speech, which also exerts a moderation effect. Thus, even in jurisdictions where platforms are immune from liability,[3] there remain rules which determine the disposition of platform discourse. By contrast, platform liability regimes, such as those embodied in Australia's Sharing of Abhorrent Violent Material Bill, Germany's Network Enforcement Act, and Singapore's Protection from Online Falsehoods and Manipulation Bill, hold platforms liable for the speech of their users. Liability requires some level of knowledge. Thus, these rules provide for platform liability when the platforms themselves fail to remove illegal content after receiving notice, but the point is that under these regimes, platforms are held liable for user speech.

Ideally, lawmakers are concerned with the institutional health of their societies. All governments, especially liberal democracies, rely on citizen discourse as a lawmaking input, and its disposition is generally understood to be positively correlated with good lawmaking and robust institutions (Rawls 1993). At a minimum, citizens that exchange views are more likely to empathize with each other and reach a broader consensus that elevates social welfare. Moreover, discourse that is markedly civil, that is, discourse in which citizens dispassionately share their views, is associated with higher levels of bargaining and preference satisfaction (Bejan 2017). Civility leads to the inclusion of social groups and engagement between opposing groups, which expands the political bargaining space and resultant gains from trade. Incivility exerts the opposite effect. While the model below does not explicitly model the effect of fake news on civility, higher levels of fake news may be conducive to higher levels of discursive conflict, disengagement, and polarization. Simultaneously, greater conflict may generate greater demand for fake news.

Regulating speech, however, is socially costly. Fake news can be difficult to ascertain. Discerning civil from incivil discourse at the margins is demanding. Thus, liberal speech regimes are generally predicated upon avoidance of error (Posner 1986). They are also predicated on the competitive screening of political ideas. Higher levels of information, enabled by loose restrictions on speech, lead to intense

[1] 52 U.S.C. § 30121 and 11 C.F.R. 110.20.

[2] For instance, the GDPR may raise barriers to entry, keep smaller platform entrants (with smaller content moderation budgets) from entering the market, and reduce overall incivility levels.

[3] See, e.g., § 230 of the U.S. Communications Decency Act.

competition and better ideas. Moreover, many of today's fora for exchanging political ideas serve to reinforce group loyalty and provide entertainment (Bloom 2016: 236). Reduction of fake news and incivility within those fora can have little impact on institutional health inasmuch as their participants are isolated. On the other hand, if a forum is considered a public good, or if private speech acts generate externalities, then there can be a basis for regulation [Packingham v. North Carolina, 582 U.S. _ (2017), Coase (1974), but see Manhattan Community Access Corp. v. Halleck, 139 S. Ct. 1921 (2019)]. As such, lawmakers interested in the institutional health of their societies should concern themselves with user-generated platform speech when the costs of market failure become excessive.

In the model below, lawmakers consider the social costs of incivil speech on platforms. By assumption, platform incivility generates institutional decay, which lawmakers seek to minimize when setting a social welfare-maximizing platform immunity policy. The model demonstrates that lawmakers prefer immunity, even if user incivility and platform non-compliance with other rules is increasing, if the costs of implementing a platform liability regime are greater than the costs of enforcing status quo law. In addition, inasmuch as implementation of a platform liability regime is coupled with speech restrictions that are unconstitutional or prohibitively costly, lawmakers prefer immunity, but platforms are free to set strong content moderation policies consistent with existing law. Thus, the private governance function of platforms highlighted by Balkin (2018), Klonick (2018), Langvardt (2018), and others is directly related to lawmakers' ability to enact and enforce alternatives, and further, it goes beyond private enforcement of existent free speech restrictions. Inasmuch as lawmakers are prohibited from suppressing unwanted speech by constitutional limits, as well as excessive lawmaking and enforcement costs, they give platforms wider discretion to make private suppression decisions. The status quo governance function of platforms, therefore, entails a private lawmaking function for determining which types of speech to suppress.

2 Model

The model describes a unit measure of platform users i and and a unit measure of platforms p, both with heterogeneous preferences independently drawn from absolutely continuous distributions. Benevolent lawmakers minimize social costs based upon their anticipation of the effect of the platform immunity regime on institutional health. This effect is referred to as incivility for exposition.

The model consists of two periods. In the first period, the actual effectiveness of the immunity policy on incivility is unknown. Lawmakers move first and choose $\lambda \in \{0, 1\}$,which is the binary decision to continue status quo immunity (0) or implement a new platform liability regime. The status quo regime has some bite. As explained above, lawmakers can enforce existing laws that impact platform content

through rules like FECA and the GDPR.[4] In addition, lawmakers can enforce existing laws that forbid illegal speech, such as defamation and incitement to imminent lawless action. Implementation of a new content moderation regime, by contrast, entails lawmaking and additional enforcement costs (together referred to as implementation costs) in order to hold platforms liable for user-generated content.

Once lawmakers choose the liability regime, platform users choose their aggregate level of incivility $\bar{k}$. Platforms then observe aggregate incivility and optimally set their internal compliance policies, which gives $\bar{h}$, the aggregate level of platform non-compliance.[5] The model permits selection of internal compliance policies on any basis, including profit, corporate image, long-term viability, good citizenship, and a desire for friendly legal environments. In the second period, lawmakers decide to continue the policy of status quo immunity or replace it with a platform liability rule. The model treats period 2 as notional. Its significance is that at the beginning of the period, lawmakers can observe a signal related to the actual effectiveness of the chosen regime in reducing incivility.

The precision of the signal, by assumption, is increasing in the level of civility (i.e. reductions in incivility) chosen by platform users, which is then reflected in platform discourse, in period 1.[6] It is further assumed that the social value of the lawmakers' decision is increasing in the precision of the signal, that is, welfare-maximizing lawmakers make better decisions as their understanding of a policy outcome increases. Thus, the underlying premise of the model is that social value of the period 2 decision is increasing for lawmakers and platform users in the period 1 level of civility of users, and its impact on platform discourse. This feature is modeled by including a "value of revealed information" term in the objective function of platform users. By backward induction, in any equilibrium, lawmakers take into account the platform users' best response and choose λ accordingly in period 1. The order of play can be summarized as follows:

1. Lawmakers choose the platform liability regime λ from set $\{0, 1\}$ where (0) is immunity and (1) provides for liability for user-generated content.
2. Platform users observe λ and choose their aggregate level of optimal incivility $\bar{k}$.
3. Platforms observe $\bar{k}$ and set internal compliance policies, which give $\bar{h}$, the aggregate level of platform non-compliance.
4. Lawmakers observe $\bar{h}$ and $\bar{k}$, and their impact on institutional health, and choose to implement liability or continue immunity.

[4] Certainly the GDPR can be understood as a response to the new threats of social media and the Internet, and thus be seen as responsive legislation. However, the GDPR does not provide for platform liability for illegal user-generated speech.

[5] Internal compliance policies are platform policies for compliance with all laws. Under a platform liability regime, these include policies for compliance with rules that hold platforms responsible for user-generated content in addition to all other rules that govern platform behavior. Under immunity, compliance policies only address rules that govern platform behavior and exclude consideration of liability for user-generated speech.

[6] This assumption is predicated upon the reasoning that the higher the level of civility, the more visible is the policy outcome.

2.1 Social media user's optimal civility

Given lawmakers' chosen platform liability regime, platform users choose their aggregate level of incivility. An individual user's utility function is given by

$$u_i(\sigma_i, \omega_i, \lambda) = -(\sigma_i, \omega_i)^2 - s_\lambda(y - \omega_i)^2 - t_\lambda(\omega_i)^2 + \beta W(\omega_i) \quad (1)$$

where σ_i is the ideal policy location for the individual i and y is the policy location of the liability regime. The policy location determines whether specific user content is sanctioned. Note that the location is flexible enough to account for any policy of content screening. The location may simply prohibit incitement to imminent lawless action, or can prohibit abhorrent violent material and fake news in addition. The location also contemplates existing restrictions on speech such as fraud, defamation, and criminal hate speech. Both σ_i and y lie anywhere on the real line. Each platform user chooses an action ω_i, such as posting or commenting on platform content, which also lies on the real line. The distance $|y - \omega_i|$ is the measure of a user's incivility, which has a proportional penalty s attached that takes the form of first-, second-, and third-party sanctions.[7] Simultaneously, complying with the policy of the liability regime can itself be costly since compliance may require changes to behavior. These costs, denoted here by the function $t_\lambda(\cdot)$, depend upon the prevailing liability regime λ, and the difference between the policy location of the existing regime, i.e. the status quo normalized to 0, and the chosen action ω_i.

The term W is realized in the second period after lawmakers observe the aggregate incivility of platform discourse and decide to continue or change the liability regime. As a result, it is discounted by the factor β, which represents the impatience of the populace for the welfare-maximizing regime. In this two-period formulation of the game, adjustment costs are incurred during the first period only so as to avoid modeling any strategic interaction in the second period. It is assumed for simplicity that lawmakers make the socially optimal decision in period 2, given the information revealed in period 1. Making this assumption avoids the need for modeling future periods repeatedly.

The function above can be rewritten with $k = y - \omega_i$, which is the distance between the policy location of the liability regime and the action chosen by the user, and, which can be interpreted as the level incivility

$$u_i(\sigma_i, \omega_i, \lambda) = -(\sigma_i, -y + k_i)^2 - s_\lambda(k_i)^2 - t_\lambda(y - k_i)^2 + \beta W(y - k_i) \quad (2)$$

[7] First-party sanctions take the form of intrapersonal guilt. Second-party sanctions take the form of interpersonal disapproval. Third-party sanctions take the form of state-imposed fines, injunctions, incarceration, expulsion, and so on. The penalty is assumed to be net of any contribution from sources of intrapersonal pride and interpersonal approval conferred by acting incivil. Thus, the model allows for a user to choose an incivil act that actually provides a net gain. In that case, the penalty would increase that user's utility, and would be better understood as a reward. The term penalty is used here for exposition given that incivility carries a negative connotation. Note, too, in jurisdictions with lax policies, the state-imposed penalty is simply 0. In that case, an incivil user immune to guilt and disapproval will experience a net penalty of 0.

Let t_λ be a quadratic cost function with fixed and marginal costs of moving away from the level of civility required by the prevailing policy of the platform liability regime

$$t_\lambda(\omega_i) = a_\lambda + b_\lambda \omega^2 = a_\lambda + b_\lambda (y - k_i)^2 \tag{3}$$

Substitute Eq. 3

$$-(\sigma_i, -y + k_i)^2 - s_\lambda (k_i)^2 - a_\lambda - b_\lambda (y - k_i)^2 + \beta W(y - k_i) \tag{4}$$

Maximizing gives the individual user's optimal level of incivility k_i

$$k_i = \frac{y - \sigma_i}{1 + s_\lambda + b_\lambda} + \frac{b_\lambda y}{1 + s_\lambda + b_\lambda} - \frac{1}{2(1 + s_\lambda + b_\lambda)} \beta W_\omega \tag{5}$$

where W_ω is the partial derivative of W with respect to ω.

As expected, incivility is higher the further the policy location of the liability regime is from the user's ideal position σ. However, the penalty s exerts a downward pressure on this response. In addition, incivility is higher the more radical the policy is, i.e. the further it is from the status quo. But again, the effect is dampened by the penalty. The marginal cost of compliance, b_λ, has a significant influence on incivility as well. Inasmuch as

$$\frac{\delta}{\delta b_\lambda} \left[\frac{b_\lambda}{1 + s_\lambda + b_\lambda} \right] > 0, \tag{6}$$

incivility is higher if the marginal cost of compliance is higher. Finally, the last term indicates that as the marginal value of information revealed by the first period level of civility increases, the lower will be incivility.

Integrating over the complete distribution of individual choices gives the aggregate k for any given distribution, among users, of ideal policy locations σ of the platform liability regime. For any given distribution of σ, e.g. $f(\sigma)$, the aggregate level of user incivility is given by

$$\bar{k} = \int k_i f(\sigma) d\sigma \tag{7}$$

Note that since integration is a linear operation, the various parameters effect $\bar{k}$ the same way they effect individual k_i.

2.2 Platform's optimal level of non-compliance

Given users' aggregate level of incivility, and platform liability regime λ, platforms set their internal content moderation policies. An individual platform's utility is given by

$$u_p(\alpha_p, \theta_p, \lambda) = -(\alpha_p, \theta_p)^2 - o_\lambda (y - \theta_p) - m_\lambda (\theta_p)^2 \tag{8}$$

where α_p is the ideal policy location of the content moderation regime for platform p and θ_p is its chosen internal compliance policy.[8] Recall that y is the policy location of the content moderation regime. Both α_p and y lie anywhere on the real line. The location of a platform's compliance policy θ_p also lies on the real line. The distance $\left|y-\theta_p\right|$ is the measure of a platform's non-compliance, which has a proportional penalty o attached in the form of fines, injunctions, legal fees, and other associated costs. Finally, administration of a content moderation policy is costly as it requires moderation of user content. These costs are denoted by the function $m_\lambda(\cdot)$ and depend upon the platform liability regime λ chosen by lawmakers, as well as the difference between the policy location of the original regime, i.e. the status quo normalized to 0, and the location of the platform's internal moderation policy θ_p.

Rewriting the equation above with $h = y - \theta_p$, which is the distance between the policy location of the platform liability regime and an internal content moderation policy chosen by platform p, and which can be interpreted as a platform's non-compliance, gives

$$u_p\left(\alpha_p, \theta_p, \lambda\right) = -\left(\alpha_p, -y + h_p\right)^2 - o_\lambda\left(h_p\right)^2 - m_\lambda\left(y - h_p\right)^2 \tag{9}$$

Let $m_\lambda(\cdot)$ be a quadratic cost function with fixed and marginal costs of moving away from the status quo

$$m_\lambda\left(\alpha_p\right) = c_\lambda + d_\lambda \alpha_p^2 = c_\lambda + d_\lambda\left(y - h_p\right)^2 \tag{10}$$

Substitute in Eq. 9 and get

$$-\left(\alpha_p, -y + h_p\right)^2 - o_\lambda\left(h_p\right)^2 - c_\lambda - d_\lambda\left(y - h_p\right)^2 \tag{11}$$

Maximizing gives the platform's optimal level of platform non-compliance h_p:

$$h_p = \frac{y - \alpha_p}{o_\lambda + d_\lambda} + \frac{d_\lambda y}{o_\lambda + d_\lambda} \tag{12}$$

Examining this expression demonstrates the trade-offs faced by the platform. First, as expected, platform non-compliance is higher the further the policy location of the platform liability regime is from the platform's preferred position.

While non-compliance is higher the greater the distance is between the platform's ideal policy location and the location set by lawmakers, the penalty o exerts a downward pressure on this response. More importantly, the second term suggests that non-compliance is higher the more radical is the new policy, i.e. the further it is from the status quo. But again, this effect is dampened by the penalty imposed.

[8] The results of the model are identical for any action θ_p chosen by the platform so long as that action reflects its choice of the location of its internal content moderation regime. For instance, θ_p can represent a chosen content moderation regime that reflects a platform's chosen level of advertising revenue, where content regimes further from y can be interpreted as advertising to larger target audiences that include users that post content prohibited by law, and those users' followers, readers, and viewers. Thus, the model permits platforms to increase profits, for instance, by increasing the space between y and θ_p.

The marginal cost of compliance, d_λ, also has significant influence on the prevailing level of platform non-compliance. Inasmuch as

$$\frac{\delta}{\delta d_\lambda}\left[\frac{d_\lambda y}{o+d_\lambda}\right] > 0, \tag{13}$$

non-compliance is higher if the marginal cost of compliance is higher.

Integrating over the whole distribution of platform non-compliance gives the aggregate h for any given distribution, among platforms, of ideal content moderation regimes. For any given distribution of α, i.e. $f(\alpha)$, the aggregate level of non-compliance across all platforms is given by

$$\bar{h} = \int h_p f(\alpha) d\alpha. \tag{14}$$

Again, as integration is a linear operation, the various parameters affect $\bar{h}$ the same way as discussed above for individual h_p.

2.3 Lawmakers' optimal content moderation policy

Given users' anticipated aggregate level of incivility $\bar{k}$ and platforms' anticipated aggregate level of non-compliance $\bar{h}$, legislators respond by minimizing the cost of institutional decay generated by incivility, i.e. maximizing the objective function

$$v\left(\lambda, \bar{k}, \bar{h}\right) = -\psi_\lambda |y| - \epsilon_\lambda \left|\bar{k}, \bar{h}\right| + \delta V\left(y - \bar{k}\right) \tag{15}$$

The first term represents the cost of enacting a new policy of holding platforms liable for user speech, which is directly proportional to the absolute distance from the status quo. By definition, any move in y must be accompanied by a switch from immunity to liability. As a result, enactment costs under immunity are 0.[9] The model can account for three possibilities when lawmakers implement platform liability for user speech. The policy location can remain the same, the policy location can increase (because, for instance, platform liability is coupled with a new rule against fake news as in Singapore's Protection from Online Falsehoods and Manipulation Bill), or the policy location can decrease as a result of, for example, a compromise or bargain struck between platforms and lawmakers.

The second term represents enforcement costs, which depend upon aggregate non-compliance $\bar{h}$ and aggregate incivility $\bar{k}$

$$\epsilon_\lambda \left|\bar{h}, \bar{k}\right| = r_\lambda \bar{h} + w_\lambda \bar{k} \tag{16}$$

By assumption, $r_1 > r_0$, since under a liability regime, platforms will be liable for failing to sufficiently moderate user speech in addition to existing rules that hold

[9] Given the current legislative proposals referenced in the introduction, the foregoing analysis seeks to evaluate moves from immunity under status quo speech screening policies to platform liability with or without changes to screening policies.

platforms liable for other reasons. On the other hand, $w_1 < w_0$ since platforms engage in greater moderation under a regime that holds them liable for user speech.

Finally, the third term represents the value of revealed information from making an optimal decision in period 2. This term is a function of aggregate user compliance, i.e. civility.

Recall that lawmakers are faced with the choice between maintaining status quo immunity and implementing a platform liability regime.They will maintain the status quo if

$$-\psi_1|y| - r_1\bar{h} - w_1\bar{k} + \delta V\left(y - \bar{k}\right) > -r_0\bar{h} - w_0\bar{k} + \delta V\left(y - \bar{k}\right) \tag{17}$$

3 Comparing welfare

Proposition 1 *Lawmakers prefer liability (immunity) as platform liability penalties increase (decrease); user penalties decrease (increase); platform marginal compliance costs decrease (increase); and user marginal compliance costs increase (decrease).*

To reduce incivility and institutional decay, lawmakers face the tradeoff between incurring platform liability enactment and enforcement costs versus the costs of enforcing existing law. Enactment costs depend upon the policy location of the speech screening policy. Enforcement costs depend upon platform non-compliance $\bar{h}$ and user incivility $\bar{k}$. Recall that aggregate platform non-compliance depends upon the distance of the policy location from the status quo, normalized to be 0; the distance of the policy location from the platforms' ideal positions $y - \alpha$, the penalties imposed for non-compliance o, and the marginal costs of compliance d. User incivility depends upon the distance of the policy location from the status quo; the distance of the policy location from the users' ideal policy positions $y - \sigma$; the penalties imposed for incivility s; and the marginal costs of compliance b.

Consider platform marginal compliance costs d. Decreased (increased) platform marginal compliance costs d decrease (increase) aggregate platform nConsider platform penalties o. Increases (decreases) in o decrease (increase) $\bar{h}$ for $\lambda = 0, 1$. However, under a liability regime, platforms will be liable for failing to sufficiently moderate user speech in addition to existing rules that hold platforms liable for other reasons, and $r_1 > r_0$ by assumption. Thus, for any o, platform enforcement costs under liability are greater than platform enforcement costs under immunity. As a result, lawmakers prefer liability (immunity) when o is increasing (decreasing) holding other factors constant.[10] Consider user penalties s. Decreases (increases) in s increase (decrease) $\bar{k}$ for $\lambda = 0, 1$. However, $w_1 < w_0$ by assumption, since platforms engage in greater moderation under a regime that holds them liable for user speech. Thus, for any s, user enforcement costs under liability are less than user enforcement costs under immunity. As a result, lawmakers prefer liability (immunity) when s is decreasing (increasing) holding other factors constant.

[10] Summary results are presented in Table 1.

Table 1 Summary results

	Platform penalties o	User penalties s	Platform marginal compliance costs d	User marginal compliance costs b	Compliance quantity via distance y	Aggregate distances from platforms' ideals $y-\alpha$	Aggregate distances from users' ideals $y-\sigma$
Platform liability enactment costs $-\psi\lvert y\rvert$	–	–	–	–	Increasing	Increasing	-
Platform enforcement costs $r\bar{h}$	Decreasing	–	Increasing	–	Increasing	Increasing	-
User enforcement costs $w\bar{k}$	–	Decreasing	–	Increasing	–	–	Increasing

Increases in quantities of the top row elements lead to corresponding ceterus paribus impacts in platform liability enactment costs, platform liability enforcement costs, and user enforcement costs

Consider platform marginal compliance costs d. Decreased (increased) platform marginal compliance costs d decrease (increase) aggregate platform non-compliance $\bar{h}$. Given that $r_1 > r_0$, lawmakers prefer liability (immunity) for decreasing (increasing) d holding other factors constant. Consider user marginal compliance costs b. Increased (decreased) user marginal compliance costs b increase (decrease) aggregate incivility $\bar{k}$. As $\bar{k}$ increases (decreases), user enforcement becomes more costly under immunity (liability) given that $w_1 < w_0$, and lawmakers prefer liability (immunity) holding other factors constant.

Proposition 2 *When moving from immunity to liability, lawmakers reduce enforcement costs the less radical is any change to the speech screening policy, and the closer any change to the speech screening policy is to the platforms' and users' ideal locations.*

By definition, lawmakers only change the location of the speech screening policy when moving from immunity to liability. Consider first, the radicalness of the change in a speech screening policy. The less radical the change, the closer the policy remains to the status quo, and as a result, less compliance costs are incurred by platforms and users, which implies smaller $\bar{h}$ and $\bar{k}$. Smaller $\bar{h}$ and $\bar{k}$ implies lower platform and user enforcement costs. Also, a less radical change implies smaller enactment costs $\psi_1|y|$, given that those costs are directly proportional to the status quo. As a result, enactment and enforcement costs under platform liability decrease as the distance between the policy location y and the status quo decreases.

Consider second, the distance of the speech screening policy from the platforms' and users' ideal locations. As the aggregate distances $y - \alpha_p$ and $y - \sigma_i$ decrease, platform non-compliance and user incivility decreases, which in turn, reduce enforcement costs under the new policy location of the platform liability regime.

Corollary 1 *Inasmuch as implementation of a platform liability regime or a move to a new speech screening policy is unconstitutional or prohibitively costly, lawmakers prefer status quo immunity, but platforms are free to set strong content moderation policies consistent with existing law.*

When implementation of a new content moderation regime is unconstitutional or prohibitively costly, $\psi_1|y| + r_1\bar{h} + w_1\bar{k} > r_0\bar{h} + w_0\bar{k}$. Lawmakers continue status quo immunity irrespective of platform moderation policies $h_p(\theta_p)$ and their influence on $\bar{h}$, and resultant impact on r.

Proposition 3 *Given a constitutionally fixed speech screening policy, lawmakers prefer platform immunity, even if user incivility is increasing, if platform enforcement costs savings under immunity exceed user enforcement cost savings under liability.*

Under status quo immunity, lawmakers are faced with the decision of implementing a liability regime given enactment costs $\psi|y|$, platform liability enforcement

costs $r\bar{h}$, and user costs $w\bar{k}$. If the speech screening policy is constitutionally fixed and remains unchanged after a move from platform immunity to liability, then enactment costs $\psi|y|$ are 0, and lawmakers are faced with $r_1\bar{h} + w_1\bar{k} > r_0\bar{h} + w_0\bar{k}$. Recall that platform enforcement costs under liability r_1 are greater than platform enforcement costs under immunity r_0 since lawmakers must enforce platform liability rules related to user speech as well as platform liability rules unrelated to user speech in addition. However, user enforcement costs under liability w_1 are less than user enforcement costs under immunity w_0 since platforms engage in greater moderation under liability. Lawmakers, therefore, prefer immunity when platform enforcement cost savings $\left(r_0 - r_1\right)\bar{h}$ under immunity exceed user enforcement cost savings $\left(w_0 - w_1\right)\bar{k}$ under liability for any level of incivility $\bar{k}$.

4 Conclusion

In many jurisdictions, platforms are immune from liability for user speech-acts. However, lawmakers in those jurisdictions may be concerned with platform civility and its impact on institutional health. In the model, lawmakers are faced with the decision to reverse a policy of platform immunity versus implementing a platform liability regime. Lawmakers prefer continued platform immunity if the costs of implementing a platform liability regime are greater than the costs of enforcing status quo law. In addition, inasmuch as implementation of a platform liability regime is coupled with new speech restrictions that are unconstitutional or prohibitively costly, lawmakers prefer immunity, but platforms are free to set strong content moderation policies consistent with existing law.

Acknowledgements The author wishes to thank two anonymous referees.

Appendix

Proof of Proposition 1 Lawmakers are faced with $-\psi_1|y| - r_1\bar{h} - w_1\bar{k} + \delta V\left(y - \bar{k}\right) > -r_0\bar{h} - w_0\bar{k} + \delta V\left(y - \bar{k}\right)$. Consider platform penalties o. Recall that platforms set optimal non-compliance so that $h_p = \frac{y-\alpha_p}{o_\lambda + d_\lambda} + \frac{d_\lambda y}{o_\lambda + d_\lambda}$. Increases (decreases) in o decrease (increase) h_p for $\lambda = 0, 1$. It follows that $\bar{h}$, given by $\int h_p f(\alpha) d\alpha$, decreases (increases) as well, for $\lambda = 0, 1$. However, $r_1 > r_0$. Thus, for any o, platform enforcement costs under liability are greater than platform enforcement costs under immunity. As a result, lawmakers prefer liability (immunity) when o is increasing (decreasing) holding other factors constant.

Lawmakers are faced with $-\psi_1|y| - r_1\bar{h} - w_1\bar{k} + \delta V\left(y - \bar{k}\right) > -r_0\bar{h} - w_0\bar{k} + \delta V\left(y - \bar{k}\right)$. Consider user penalties s. Recall that users set optimal incivility so that $k_i = \frac{y-\sigma_i}{1+s_\lambda+b_\lambda} + \frac{b_\lambda y}{1+s_\lambda+b_\lambda} - \frac{1}{2\left(1+s_\lambda+b_\lambda\right)}\beta W_\omega$. Decreases (increases) in s increase (decrease) k_i for $\lambda = 0, 1$. It follows that $\bar{k}$, given by $\bar{k} = \int k_i f(\sigma) d\sigma$, increases (decreases) as well, for $\lambda = 0, 1$. However, $w_1 < w_0$. Thus, for any s, user

enforcement costs under liability are less than user enforcement costs under immunity. As a result, lawmakers prefer liability (immunity) when s is decreasing (increasing) holding other factors constant. □

Proof of Proposition 2 When moving from immunity to liability, lawmakers incur costs $\psi_1|y| + r_1\bar{h} + w_1\bar{k}$. Consider first, the radicalness of the change to the speech screening policy. Enactment costs ψ are directly proportional to the status quo, normalized to be 0. Smaller distances from the status quo $y - 0$ result in smaller ψ.

Consider second, the distances between the new policy location and the platforms' and users' ideal policy locations, $y - \alpha_p$ and $y - \sigma_i$, respectively. Recall that platforms set optimal non-compliance so that $h_p = \frac{y-\alpha_p}{o_\lambda+d_\lambda} + \frac{d_\lambda y}{o_\lambda+d_\lambda}$ and that users set optimal incivility so that $k_i = \frac{y-\sigma_i}{1+s_\lambda+b_\lambda} + \frac{b_\lambda y}{1+s_\lambda+b_\lambda} - \frac{1}{2(1+s_\lambda+b_\lambda)}\beta W_\omega$. Decreases in $y - \alpha_p$ and $y - \sigma_i$ reduce h_p and k_i, respectively. It follows that $\bar{h}$ and $\bar{k}$ decrease, and that lawmakers incur fewer enforcement costs r and w. □

Proof of Proposition 3 Given a constitutionally fixed speech screening policy, lawmakers cannot change y. Enactment costs $\psi_1|y|$, given by $y - 0$, equal 0. Lawmakers, therefore, are faced with $r_1\bar{h} + w_1\bar{k} > r_0\bar{h} + w_0\bar{k}$, where $r_1 > r_0$ and $w_1 < w_0$, and prefer immunity when $(r_0 - r_1)\bar{h} > (w_0 - w_1)\bar{k}$ for any level of aggregate incivility $\bar{k}$. □

References

Balkin, J. (2018). Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UC Davis Law Review*, *51*, 1149–1210.

Bejan, T. M. (2017). *Mere civility: Disagreement and the limits of toleration*. Cambridge: Harvard University Press.

Bloom, P. (2016). *Against empathy: The case for rational compassion*. New York: HarperCollins.

Coase, R. (1974). The market for goods and the market for ideas. *American Economic Review*, *64*, 384–391.

Fagan, F. (2018). Systemic social media regulation. *Duke Law and Technology Review*, *16*, 393–439.

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, *131*, 1598–1670.

Langvardt, K. (2018). Regulating online content moderation. *Georgetown Law Journal*, *106*, 1353–1388.

Posner, R. A. (1986). Free speech in an economic perspective. *Suffolk University Law Review*, *20*, 1.

Rawls, J. (1993). *Political liberalism*. New York, NY: Columbia University Press.