The Role of Shaming as a Regulatory Tool:
Re-thinking Truth as an Absolute Bar to Defamation♠

Tomer Blumkin* , Yoram Margalioth♥ & Itamar Levin Fridman♣

Abstract

Contemporary defamation law fails to account for the growing social harm caused by truth-based shaming, which, while perhaps justifiable in the past, is no longer defensible in the age of social networks. This paper re-examines libel law through the analytical framework of public economics, focusing on the maximization of social welfare. Individuals act as both potential violators and enforcers of social norms, and their shaming behavior can be understood as the private provision of a public good that generates both positive and negative externalities. Using a formal model, we show that the unregulated "market" for shaming is generally inefficient: it tends to either under- or over-provide punishment relative to the social optimum. We further demonstrate that libel law can play a corrective, Pigouvian, role, by internalizing the externalities associated with shaming, even when the underlying statements are true. We challenge the prevailing U.S. doctrine that truth operates as an absolute bar to defamation liability. Our findings suggest that this rule is socially suboptimal, providing an economic rationale for re-examining defamation law in light of contemporary information dynamics and social welfare considerations.

1. Introduction

Public shaming, through media exposure, viral posts, or other digital amplification—has become a central mechanism of informal norm enforcement. When initiated by private individuals, shaming operates outside institutional boundaries, yet it powerfully shapes reputations and behavior. The First Amendment protects expressive acts of this kind whether the statements are true or false (*U.S. v. Alvarez* 2012). In defamation doctrine, courts balance freedom of expression against reputational interests only when the statement is false. Even false statements about public officials remain protected absent proof of "actual malice," that is, knowledge of falsity or reckless disregard for truth (*N.Y. Times Co. v. Sullivan* 1964). The result is categorical: truth functions as

♠ The paper is partially based on a thesis written by the third author, Itamar, under the mentorship of the first author, and is dedicated to Itamar, who died prematurely at the age of 34, leaving after him a loving and proud family.

* Department of Economics, Ben-Gurion University of the Negev, Beer-Sheba 84105, Israel, CesIfo, IZA. E-mail: tomerblu@exchange.bgu.ac.il

♥ The Buchmann Faculty of Law, Tel-Aviv University, Tel-Aviv 69978, Israel. E-mail: margalio@tauex.tau.ac.il

♣ Department of Economics, Ben-Gurion University of the Negev, Beer-Sheba 84105, Israel. E-mail: tomerblu@exchange.bgu.ac.il

an absolute defense, precluding liability regardless of the harm inflicted (*Restatement [Second] of Torts* § 581A).

This doctrine reflects the prevailing economic and legal assumption that truthful speech generates only positive externalities. Information based on accurate facts is conceived as a public good the disclosure of which promotes efficient social coordination; accordingly, the law should minimize the risk of chilling its dissemination (Farber 1991; Hylton 1996; Posner 1997; Garoupa 1999a, 1999b; Cooter 2000). Economic analyses of defamation therefore focus almost exclusively on false statements, leaving harms arising from truthful exposure largely unexamined. The sole early exception, Bar-Gill and Hamdani (2002, 2003), acknowledged that truth can also generate injury but retained the premise that veracity should bar recovery. Subsequent contributions on audience and credibility effects (Hemel and Porat 2019; Arbel and Mungan 2019, 2023; Arbel 2023) center on the dynamics of falsehood and perception rather than the welfare implications of truth-based shaming, implicitly leaving that question open.

We argue that this assumption is untenable in the digital age. As Sunstein (2020) observed, *Sullivan* "might as well have been decided a century ago" given the technological transformation of communication. Social media enables instantaneous and global dissemination of reputational sanctions, magnifying both their reach and severity (Ronson 2015; Klonick 2016). Unlike formal punishment, online shaming lacks due process, proportionality, and mechanisms of rehabilitation. Minor misconduct—or even socially ambiguous behavior—can trigger viral condemnation whose effects persist indefinitely. Existing U.S. torts, such as public disclosure of private facts or intentional infliction of emotional distress, provide little remedy: the broad "newsworthiness" defense and the high "outrageousness" threshold effectively immunize the dissemination of true information. The result is a regime of structural impunity for reputational harm inflicted through accurate but punitive speech.

The American Law Institute's ongoing *Restatement (Third) of Torts: Defamation and Privacy* (2019) reflects growing recognition that digital dissemination challenges the premises of traditional doctrine. The ALI itself has noted that the revision aims "to cover new issues, particularly those related to the Internet." We contribute to this reconsideration by analyzing the welfare consequences of truth-based shaming and offering an efficiency-oriented framework for potential reform.

This American exceptionalism becomes clearer when contrasted with the European approach to data protection and reputational privacy. Comparative analysis underscores the distinctiveness of the U.S. system and its failure to adapt to contemporary information and communication technologies. European data-protection and human-rights law treat the continued dissemination of truthful personal information as potentially unlawful when it becomes disproportionate to any legitimate public interest. Beyond the General Data Protection Regulation (GDPR) (Regulation [EU] 2016/679), Articles 7 and 8 of the Charter of Fundamental Rights of the European Union and Article 8 of the European Convention on Human Rights recognize autonomous rights to privacy and data protection. The GDPR operationalizes these rights through principles of purpose limitation and proportionality: personal data must be collected for specified, legitimate purposes and not further processed incompatibly (Art. 5(1)(b)), must be adequate and not excessive (Art. 5(1)(c)), and must be retained only as long as necessary (Art. 5(1)(e)). Even accurate data may therefore become unlawful to process when it is outdated or irrelevant. Article 17 codifies this logic in the "right to erasure," empowering individuals to demand deletion or de-indexing of personal data that are obsolete or excessive. In *Google Spain SL v. AEPD and González* (CJEU 2014, Case C-131/12), the Court of Justice of the European Union required removal of truthful but obsolete information from search results. Similarly, the European Court of Human Rights—through *Von Hannover v. Germany* (Eur. Ct. H.R. 2004, 2012) and *M.L. and W.W. v. Germany* (Eur. Ct. H.R. 2018)—has held that truthful publication may violate privacy when disclosure is disproportionate to its public-interest value. Taken together with the Council of Europe's modernized *Convention 108+* (Council of Europe 2018), these instruments establish a coherent regime of informational self-determination in which the truth of information does not immunize its circulation from legal constraint.

Our analysis proceeds from a public-economics perspective. Freedom of expression can be viewed as a market for information transmission, and shaming as a form of privately supplied enforcement of social norms. Individuals act as both potential violators and enforcers, motivated by intrinsic "warm-glow" benefits (Andreoni 1987, 1990). This market exhibits both positive and negative externalities: while exposure deters misconduct, it also imposes stigma whose social costs may exceed its benefits. We construct a formal model to identify the socially optimal level of shaming and the conditions under which unregulated, truth-based sanctions become excessive. The model provides a normative foundation for calibrating legal intervention to internalize these externalities.

The remainder of the paper is organized as follows. Section 2 develops the analytical framework. Section 3 derives the welfare implications of shaming. Section 4 explores the regulatory role of libel suits as Pigouvian instruments. Section 5 defies the prevailing doctrine viewing truth as as an absolute defense against defamation. Section 6 extends the model to heterogeneous information and herding effects. Section 7 concludes.

2. An Analytical Framework

We present a parsimonious setup which captures the key ingredients necessary for our analysis. We begin by considering a benchmark with no shaming in place and then introduce the shaming mechanism.

2.1 A Benchmark Setup with no Shaming

Consider a community with a unit measure of risk-neutral agents. Each agent has an endowment of $y>0$. Agents choose whether to violate a social norm and differ in the cost incurred in doing so, denoted by θ, where θ is i.i.d. and is assumed to be uniformly distributed over the support [0,1]. θ reflects both the monetary and psychic costs entailed by engaging in misconduct. Focusing on the role of shaming, we simplify the exposition by not modeling explicitly alternative enforcement mechanisms (which may hence be reflected in the value of θ). Denoting the (uniform) benefit from engaging in misconduct by $0<b<1$, a type-θ agent engages in misconduct if-and-only-if the following condition holds:

(1) $b \geq \theta$.

In the absence of shaming, the level of misconduct is determined by a threshold $\hat{\theta} = b$, such that agents with $\theta \leq \hat{\theta}$ choose to engage, whereas all other agents (with levels of θ exceeding the threshold) abide by the norms. By virtue of our parametric assumptions, $0 < \hat{\theta} < 1$ denotes the level of misconduct in community.

2.2 Social Stigma and Shaming

An individual involved in misconduct incurs social stigma through exposure within the community's social network (shaming). Each member of the community receives a noisy signal, z, about every other agent in the community, which is based on whether this agent has engaged in misconduct.

Formally, the probability distribution of the signal associated with agent *i* received by agent *j* takes a simple form:

(2)

$$Pr\left[z^{ij}\middle|c^{i}\right]=\begin{Bmatrix} q & z^{ij}=h;c^{i}=1 \\ (1-q) & z^{ij}=h;c^{i}=0 \\ (1-q) & z^{ij}=l;c^{i}=1 \\ q & z^{ij}=l;c^{i}=0 \end{Bmatrix},$$

where $1/2<q<1$, $h>l>0$, and c^{i} is an indicator function obtaining the value of 1 if agent *i* commits an act of misconduct and 0 otherwise.

Several observations are in order. First, notice that the signal is informative, as the likelihood of obtaining a high (low) value of the signal is increasing (decreasing) when the agent commits an act of misconduct as $q>1/2$. Second, notice that as $q<1$, the signal is noisy, hence receiving high/low levels of the signal occurs with a positive probability regardless of whether agent *i* commits an act of misconduct (with $q=1$, the signal is fully revealing and enables perfect screening). Finally, notice that signals are assumed for simplicity to be independently distributed across members of the community.

Based on the signal drawn, an agent decides whether to share the information with the other members of the community. The agent derives some benefit from exposure via the social network (through sharing his private information with others) but also entails some reputational cost from sharing false information. We simplify by assuming that the agent only shares 'bad' news, when receiving a high signal. The gain from sharing information may reflect an intrinsic benefit from exposing misconduct. In case the benefit from exposure exceeds the reputational cost, the agent is sharing the information by posting a comment (e.g., posting a 'dislike' gesture) on the public domain of the social network. Otherwise, when either the cost outweighs the benefit or when

receiving a low signal, the agent is not sharing any information by assumption.[1] [2] Formally, by virtue of earlier assumptions, applying Bayes' Rule, the probability of sharing false news, namely, assuming a 'dislike' has been posted, the probability that an agent has not been engaged in misconduct conditional on observing a high signal, is given by:

(3) $\hat{p} \equiv \frac{(1-q)\cdot(1-\hat{\theta})}{(1-q)\cdot(1-\hat{\theta})+q\cdot\hat{\theta}}$.

Let $0<F<1$ denote the reputational cost and $0 \leq K \leq 1$ denote the benefit from exposure. An agent obtaining a high signal is posting a 'dislike' if-and-only-if the following condition holds:

(4) $K > \hat{p} \cdot F$.

Assuming agents differ in the benefit derived from exposure, and further letting K be independently drawn from a uniform distribution over the support [0,1], the probability of posting a 'dislike' conditional on receiving a high signal is given by:

(5) $h \equiv [1 - \hat{p} \cdot F]$.

For each profile of realized signals associated with agent i, let n^i denote the number of 'dislikes' posted on the social network. We assume that the stigma cost entailed by an agent who is perceived to be engaged in misconduct is given by a strictly increasing function of the number of 'dislikes' posted on the social network (for tractability we assume a linear functional form).[3] Formally, the social stigma entailed by individual i engaged in misconduct is given by:

(6) $S^i_{c=1} = n^i_{c=1} \cdot S = q \cdot h \cdot S$,

whereas the social stigma entailed by individual i who abides by the norm is given by:

[1] Allowing agents to share 'good' news, by posting a 'like' gesture on the public domain, would complicate the derivations but would have no impact on the qualitative nature of the results.

[2] We simplify by assuming that agents do not directly engage in validation (fact-checking) prior to releasing the information. Assuming the latter would complicate the analysis but would not change the qualitative nature of our results. The latter follows from to a key feature of our model: the presence of reputational costs associated with sharing inaccurate information. This implies that agents are more likely to share accurate information, essentially engaging in validation. The possibility to engage in (costly) fact-checking would enhance the accuracy of the information shared on the public domain. In the presence of validation, the government may consider subsidizing fact-checking and certifying validated shared information, on Pigouvian grounds, to internalize the informational spillover.

[3] If agents share 'good' news (realization of low signals) stigma may be captured by a strictly increasing function of the difference between the number of posted 'dis-likes' and the number of posted 'likes'.

(7) $S^i_{c=0} = n^i_{c=0} \cdot S = (1-q) \cdot h \cdot S,$

where $0<S<b$ denotes the stigma cost, $n^i_{c=1}$ denotes the number of 'dislikes' posted when *i* has been engaged in misconduct and $n^i_{c=0}$ denotes the number of 'dislikes' posted when *i* abides by the social norms.

Notice that *S*, which denotes the stigma cost, reflects the disutility from shaming entailed by agents perceived to be violating the social norms. The latter may be broadly interpreted to encompass both psychic and pecuniary costs, that are endogenously determined in equilibrium and reflect the community response to the exposure of misconduct. The level of stigma would be affected by the number of community members who were exposed to the tarnishing information and interact with the shamed individual. Members of the community may choose to limit their social interactions or avoid trading with agents who are exposed to be engaged in misconduct [Arbel and Mungan (2023) refer to these costs as stemming from an 'audience effect']. This effect would be compounded by the length of time during which the tarnishing information is available to the public. For tractability we assume that *S* is an exogenous parameter.

Comparing the expressions in (6) and (7), it follows, as $q>1/2$, that $S^i_{c=1} > S^i_{c=0} > 0$. Namely, the stigma entailed by an agent committing an act of misconduct strictly exceeds that entailed by a norm-abiding agent, which implies that shaming may potentially serve for screening/self-enforcement purposes.[4]

In the presence of shaming, a type-θ agent commits an act of misconduct if-and-only-if the following condition holds:

(8) $b - [S_{c=1} - S_{c=0}] \geq \theta.$

Clearly, the presence of stigma serves to enhance deterrence and thereby reducing the level of misconduct, determined by a cutoff rule:

(9) $\hat{\theta} = b - [S_{c=1} - S_{c=0}].$

[4] Notice that our focus is on the role of shaming in reducing the level of misconduct via enhanced deterrence. There is clearly an additional, presumably significant, role of shaming in preventing such undesirable acts (incapacitation), by providing information to the public about potential risks (say, the proximity of agents with known predisposition to engage in sexual harassment). Further notice that even when the agent is norm-abiding, a strictly positive level of stigma is, nevertheless, entailed as the signal is noisy.

Substituting from (6) and (7) for the social stigma terms yields:

(10) $\hat{\theta} = b - (2q-1) \cdot h \cdot S.$

Substituting for $\hat{p}$ from (3) into (5) yields:

(11) $h = 1 - \frac{(1-q)\cdot(1-\hat{\theta})}{(1-q)\cdot(1-\hat{\theta})+q\cdot\hat{\theta}} \cdot F.$

The equilibrium in the presence of shaming is given by a solution of the system of two equations [(10) and (11)] for h and $\hat{\theta}$, where h measures the extent of shaming and $\hat{\theta}$ denotes the level of misconduct. It is straightforward to verify existence and uniqueness of the equilibrium, in which $0 < \hat{\theta} < 1$ and $0<h<1$.

3. The Socially Optimal Level of Shaming

In the current section we turn to evaluate the welfare implications of shaming by comparing the *laissez-faire* market allocation (absent of intervention) with the socially optimal allocation. We assume a utilitarian social planner that aims to maximize welfare by choosing the extent of shaming, h. Formally,

(12) $max_h W(h) \equiv$

$$\beta \cdot \int_0^{\hat{\theta}(h)} \Big[y + b - \theta - q \cdot h \cdot S + \gamma \cdot \big[(1-q)\cdot\big(1-\hat{\theta}(h)\big) + q\cdot\hat{\theta}(h)\big] \cdot \int_{1-h}^{1} \big[K - \hat{p}[\hat{\theta}(h)]\cdot F\big]dK - \frac{\alpha}{2}\cdot[\hat{\theta}(h)]^2 \Big] d\theta$$

$$+[1-\beta\cdot\hat{\theta}(h)] \cdot \Big[y - (1-q)\cdot h\cdot S + \gamma\cdot\Big[(1-q)\cdot\Big(1-\hat{\theta}(h)\Big) + q\cdot\hat{\theta}(h)\Big] \cdot \int_{1-h}^{1} \big[K - \hat{p}\big[\hat{\theta}(h)\big]\cdot F\big]dK \quad - \frac{\alpha}{2}\cdot[\hat{\theta}(h)]^2 \Big]$$

where $\hat{\theta}(h) = b - (2q-1)\cdot h\cdot S$ denotes the level of misconduct, $\frac{\alpha}{2}\cdot\big[\hat{\theta}(h)\big]^2$ denotes the social cost of misconduct (assumed to be quadratic for tractability) with $\alpha > 0$ measuring the severity of

the misconduct in the eyes of the community members; $0 \leq \beta \leq 1$ denotes the relative social weight assigned to community members who do not abide by the social norms and $0 \leq \gamma \leq 1$ denotes the social weight assigned to "warm glow" (Andreoni 1987, 1990) associated with shaming, that is, the intrinsic utility derived from sharing information about acts of misconduct with other members of the community.

Several observations are worth noting. First, notice that we model misconduct as a general public good ('bad') shared by all members of the community. Second, notice that shaming and misconduct generate negative externalities not internalized by the agents, and hence constitute a source of potential inefficiency, which may warrant qualifying the freedom of expression of community members (regulating the market for information transmission). Notably, there is an inherent trade-off between the two, as an increase in the extent of shaming serves to reduce the level of misconduct. Notice that we allow for an asymmetric treatment of individuals who engage in misconduct and norm-abiding agents in the social calculus. Notably, when $\beta < 1$ the social planner's objective function captures a 're-distributive' motive in favor of the norm-abiding individuals. Finally, notice that as is common in the literature on the private provision of public goods [see e.g, Diamond (2006)], by considering the case where $\gamma < 1$, we allow for the possibility to 'launder-out' (wholly or partially) the benefit from 'warm glow' from the social calculus, due to potential 'double counting' of the benefit generated by the public good in the social welfare function. In our context, individuals contribute to the reduction of misconduct (a public good shared by the entire community) by engaging in shaming, from which they also derive intrinsic benefits. Thus, the benefit from the provision of the public good (enhanced enforcement of norms) appears twice in the welfare function. The presence of 'warm glow' (if accounted for in the social calculus) may provide a normative justification for the private provision of shaming (in contrast to government provision).

To simplify our exposition, we will henceforth drop the 'warm glow' component from the social welfare function by setting $\gamma = 0$.[5] In this case the maximization in (12) simplifies to:

[5] Notice, that incorporating the 'warm glow' component into the welfare function (that is, setting $\gamma > 0$) will introduce an additional negative externality due to shaming. An increase in the extent of shaming, h, will induce a decrease in the extent of misconduct, $\hat{\theta}$, which in turn will reduce the likelihood of being engaged in shaming (the probability of observing a 'bad' signal) and will also increase the reputational cost entailed by shamers (due to false shaming). Both effects contribute to a decrease in the expected (net) 'warm-glow' gains from shaming.

(12') $max_h W(h) \equiv$

$$\beta \cdot \int_0^{\hat{\theta}(h)} [y + b - \theta - q \cdot h \cdot S] d\theta + \left[1 - \beta \cdot \hat{\theta}(h)\right] \cdot [y - (1-q) \cdot h \cdot S] - \frac{\alpha}{2} \cdot [\hat{\theta}(h)]^2,$$

where $\hat{\theta}(h) = b - (2q-1) \cdot h \cdot S$.

Due to the presence of externalities (barring a knife-edge case in which the two externalities precisely offset one another under *laissez-faire*) it is evident that the market allocation is sub-optimal from the social planner's perspective and intervention is warranted. To demonstrate this, we evaluate the derivative of welfare with respect to h at the 'laissez-faire' level of shaming. Formally, let $0 < h^* < 1$ denote that 'laissez-faire' level of shaming given by the solution to the system of two equations (10) and (11). Differentiating the welfare function with respect to h and evaluating the derivative at $h=h^*$ yields, upon re-arrangement:

(13) $\left.\frac{\partial W}{\partial h}\right|_{h=h^*} =$

$$-\left[\beta \cdot \hat{\theta}(h^*) \cdot q \cdot S + \left(1 - \beta \cdot \hat{\theta}(h^*)\right) \cdot (1-q) \cdot S\right] + \left[\alpha \cdot \hat{\theta}(h^*) \cdot (2q-1) \cdot S\right]$$

The expression on the right-hand side of (13) reflects the trade-off between the gain from shaming (due to the induced reduction in the level of misconduct), captured by the second (positive) term on the right-hand side of (13); and, the cost of shaming due to the stigmatization entailed both by members of the community who engage in misconduct and those abiding by the social norms, captured by the first (negative) term on the right-hand side of (13). Signing the expression is generally ambiguous and depends on the parametric assumptions. Barring knife-edge cases, the sign of the derivative is non-zero, hence the level of shaming under 'laissez-faire' is suboptimal.

The optimal level of shaming depends on the parametric assumptions being invoked. To gain some intuition we will consider several special cases. When q=1/2 the signal is uninformative. In this case, shaming does not serve to enhance enforcement ($\hat{\theta} = b$, regardless of h), but does entail stigma costs, hence the sign of the derivative of the welfare expression in (12') with respect to h will be unambiguously negative for any h. Social-optimum will be given by a corner solution, h=0, namely shaming should be ruled out. When q=1, the signal is perfectly informative. In this case,

shaming is perfectly targeted towards individuals who engage in misconduct and no stigma is entailed by norm-abiding agents. Prima-facie, this seems to be the strongest case for shaming as an enforcement tool. Differentiating the welfare expression in (12') with respect to h yields after re-arrangement:

$$(14) \quad \left.\frac{\partial W}{\partial h}\right|_{q=1} = (\alpha - \beta) \cdot \hat{\theta}(h) \cdot S.$$

The social optimum again is obtained as a corner solution but depends on the parametric assumptions. If $\alpha > \beta$, the sign of the derivative in (14) is unambiguously positive for any h. In this case, the social gain from the reduction of misconduct is sufficiently high relative to the stigma entailed by individuals who engage in misconduct; hence, we obtain the anticipated result - social optimum is given by h=1, that is setting shaming to the maximum extent feasible.

If $\alpha < \beta$, however, the sign of the derivative in (14) is unambiguously negative for any h, hence, social optimum is given by h=0, that is shaming should be ruled out. This result is somewhat surprising as, in this case, shaming is based on perfectly accurate evidence. However, as members of the community who engage in misconduct are not fully laundered out from the social calculus ($\beta > 0$), one should account for the significant shaming they are exposed to via the social network. If the social gain from enhanced enforcement of norms via shaming is moderate, shaming turns out to be welfare detrimental. Notably, the decision whether to engage in shaming does not reflect the severity of the act of misconduct (K , the benefit from exposure, is by assumption independent of α). This strong assumption captures in sharpest relief a plausible assumption about the 'non-proportionality' of social shaming: the existence of a potentially positive but rather weak correlation between the private incentives to engage in shaming and the severity of the underlying act of misconduct. Indeed, shaming is often criticized on the grounds of being an excessive tool to address acts of misconduct. Individuals may be deterred from engaging in acts deemed socially undesirable if subject to shaming, but the cost entailed by those who do engage in such activities may be too harsh (ostracism, lost career etc.). This may happen even in circumstances where shaming is based on well-established evidence, which seems to provide the strongest case in its support.

One could in principle allow for the 'warm glow' benefit from shaming to be positively related with the severity of the act of misconduct, by allowing K to be an increasing function of α.

However, if the correlation between the private gains from shaming (hence, the propensity to engage in shaming) and the severity of the act of misconduct (subject to shaming) is sufficiently weak, the market equilibrium will typically yield an excessive level of shaming in cases where the social value of enhanced enforcement via shaming is relatively small (α is small). Another factor which contributes to the weak correlation alluded to above is the presence of herding pattens commonly documented in social networks. Herding can be simply captured by allowing K to be an increasing function of the extent of shaming, h. That is, the individual's propensity to engage in shaming is increasing in the prevalence of shaming. The latter would imply that even when the intrinsic benefits from shaming would be small (say, due to a small value of α, reflecting limited public interest in the act of misconduct and/or in the individual who has committed the act) the level of shaming could potentially explode due to the presence of a strong herding effect. We introduce herding effects and discuss their implications in Section 5 below.

The corner solutions obtained under the scenarios analyzed above clearly extend by continuity to the respective cases where q is either sufficiently close to ½ or sufficiently close to 1. An interior solution, where $0<h<1$, may be obtained when $\alpha > \beta$, $½<q<1$ and q is bounded away from 1 and ½. In this case the social optimum will satisfy the following first order condition:

$$\frac{\partial W}{\partial h} = -\left[\beta \cdot \hat{\theta}(h) \cdot q \cdot S + \left(1 - \beta \cdot \hat{\theta}(h)\right) \cdot (1-q) \cdot S\right] + \left[\alpha \cdot \hat{\theta}(h) \cdot (2q-1) \cdot S\right] = 0 \,.$$

One can verify that the second-order condition is satisfied as $\alpha > \beta$. The optimal level of shaming can be shown to increase in α and decrease in β and S, as anticipated.

4. Libel Suits as a Regulating Device

4.1 The Market Equilibrium in the Presence of Libel Suits

In light of the observation that the market equilibrium generally yields a suboptimal allocation, often generating an excessive level of shaming, in the current section we turn to examine the potentially welfare-enhancing role of libel-suits submitted by individual*s* who were subject to shaming as a regulating device.

Bringing the case to court, by submitting a libel suit, we assume that the probability that the information will be proven to be false depends on the type of the individual and for simplicity is

assumed to be given by θ, the plaintiff's cost incurred if committing an act of misconduct. Thus, individuals with a lower propensity to engage in misconduct reflected in a higher value of θ and who are subject to shaming via the social network are, plausibly, more likely to win a libel suit. Denote by d>0, the level of damages awarded to the party winning in the libel suit (plaintiff or defendant). The expected payoff associated with filing a libel suit by type- θ individual conditional on being subject to shaming is hence given by:

(15) $\quad v(d,\theta) = \theta \cdot d - (1-\theta) \cdot d = (2\theta - 1) \cdot d \geq 0 \leftrightarrow \theta \geq 1/2$.

Thus, an individual is filing a libel suit if-and-only-if $\theta \geq 1/2$.[6] Recalling that the threshold for engaging in misconduct is given by $\hat{\theta}$, it follows that when $\hat{\theta} > 1/2$, a positive fraction of the population that engages in misconduct finds it desirable to file a libel suit. Plausibly, libel suits are not excluded to norm-abiding individuals. The presence of libel suits impacts both the decision whether to commit an act of misconduct and the decision whether to engage in shaming. We start by looking at the decision problem faced by an individual who decides whether to engage in misconduct. A θ-type individual is committing an act of misconduct if-and-only-if the following condition holds:

(16) $\quad b - (2q-1) \cdot h \cdot S + (2q-1) \cdot h \cdot \max\left[0, (2\theta - 1) \cdot d\right] \geq \theta.$

The third term on the left-hand-side expression of the inequality condition in (16) captures the difference between the expected payoff from filing libel suits when committing an act of misconduct and the expected payoff from doing so when abiding by the norms.

Assuming that $2 \cdot (2q-1) \cdot h \cdot d < 1$, the level of misconduct is determined by a threshold, $\hat{\theta}$, satisfying the condition in (16) as an equality:

(17) $\quad \hat{\theta} = b - (2q-1) \cdot h \cdot \left[S - \max\left[0, \left(2\hat{\theta} - 1\right) \cdot d\right]\right].$

[6] We simplify the exposition by assuming no transaction costs and by abstracting from introducing endowment constraints. Incorporating transaction costs such as those incurred in filing a suit (which are not reimbursed if the case is resolved in favor of the plaintiff, as in the US system) or introducing the possibility of judgment-proofness, would change the cutoff. The former would increase the cutoff above ½, whereas the latter would reduce it below ½, without changing the qualitative nature of our results. Notice further that if we alternatively assume that the court can ex-post verify, with probability 1, whether the plaintiff has been engaged in misconduct, then the threshold would be given by $\hat{\theta}$, as in this scenario only those who abide by the norms and were subject to shaming will find it optimal to file a libel suit (for further implications of this alternative assumption see footnote 7 below).

It follows that for a given extent of shaming, h, the possibility to file libel suits is mitigating the deterrence effect of shaming and thereby contributes to enhanced misconduct. Notice that if $\hat{\theta}$, given by the implicit solution to (17), is (weakly) lower than ½, no such direct mitigating effect exists (all agents who engage in misconduct find it undesirable to file a libel suit). However, libel suits can still indirectly impact the level of misconduct via their effect on the extent of shaming.

Consider individual *j* who receives a high ('bad') signal about individual *i*. Individual *j* needs to decide whether to share the information over the network by posting a 'dislike'. Sharing the information subjects individual *j* to the risk of a libel suit filed by individual *i*. Notice, that individual *j* only observes the signal and does not know whether *i* has committed an act of misconduct. From the point of view of individual *j*, applying Bayes' Rule, conditional on observing a high signal about *i*, the probability that individual *i* would file a libel suit and prove that the information being shared by *j* is false/true is given, respectively, by:

$$(18) \quad g^{+} = \left\{ \frac{\left[\max\left(\hat{\theta},\frac{1}{2}\right)-\frac{1}{2}\right]\cdot q}{\hat{\theta}\cdot q+(1-\hat{\theta})\cdot(1-q)} \cdot \frac{\left(\hat{\theta}+\frac{1}{2}\right)}{2} \right\} + \left\{ \frac{\left[1-\max\left(\hat{\theta},\frac{1}{2}\right)\right]\cdot(1-q)}{\hat{\theta}\cdot q+(1-\hat{\theta})\cdot(1-q)} \cdot \frac{\left[1+\max\left(\hat{\theta},\frac{1}{2}\right)\right]}{2} \right\},$$

$$(19) \quad g^{-} = \left\{ \frac{\left[\max\left(\hat{\theta},\frac{1}{2}\right)-\frac{1}{2}\right]\cdot q}{\hat{\theta}\cdot q+(1-\hat{\theta})\cdot(1-q)} \cdot \left[1 - \frac{\left(\hat{\theta}+\frac{1}{2}\right)}{2}\right] \right\} + \left\{ \frac{\left[1-\max\left(\hat{\theta},\frac{1}{2}\right)\right]\cdot(1-q)}{\hat{\theta}\cdot q+(1-\hat{\theta})\cdot(1-q)} \cdot \left[1 - \frac{\left[1+\max\left(\hat{\theta},\frac{1}{2}\right)\right]}{2}\right] \right\}.$$

The superscripts "+" and "-" refer, respectively, to "success" and "failure", in the resolution of the libel suit (from the point of view of the plaintiff). Invoking our earlier assumptions, an agent obtaining a high signal is posting a 'dislike' if-and-only-if the following condition holds:

$$(20) \quad K > \hat{p} \cdot F + (g^{+} - g^{-}) \cdot d.$$

The probability of posting a 'dislike' conditional on receiving a high signal is hence given by:

$$(21) \quad h \equiv [1 - \hat{p} \cdot F - (g^{+} - g^{-}) \cdot d].$$

Notice that as, by virtue of our parametric assumptions, $g^{+} > g^{-}$, the presence of libel suits serves to reduce the extent of shaming. Substituting for g^{+} and g^{-} from (18) and (19) into (21) and re-arranging yields:

$$(22) \quad h \equiv 1 - \frac{(1-q)\cdot(1-\hat{\theta})}{(1-q)\cdot(1-\hat{\theta})+q\cdot\hat{\theta}} \cdot F$$

$$-\left\{\frac{\left[\max\left(\hat{\theta},1/2\right)-1/2\right]\cdot q}{\hat{\theta}\cdot q+(1-\hat{\theta})\cdot(1-q)}\cdot\left(\hat{\theta}-1/2\right)+\frac{\left[1-\max\left(\hat{\theta},1/2\right)\right]\cdot(1-q)}{\hat{\theta}\cdot q+(1-\hat{\theta})\cdot(1-q)}\cdot\max\left(\hat{\theta},1/2\right)\right\}\cdot d.$$

The equilibrium in the presence of shaming and libel suits is given by a solution of the system of two equations [(17) and (22)] for h and $\hat{\theta}$. Let $\hat{\theta}(d)$ and $h(d)$ denote the solution for the system of equations (17) and (22) as a function of d.[7] In the *laissez-faire* allocation d=0 and the equilibrium exists and is unique. By continuity these properties extend to d>0 sufficiently close to zero.

4.2 The Social Desirability of Libel Suits

We assume that the social planner aims to maximize welfare by choosing the level of libel damages, d. We invoke the same welfare specification as in (12') but simplify the exposition by letting $\beta = 1$. Notice that in the case where $\beta < 1$, libel damages play a re-distributive role. Plaintiffs are more likely to be norm-abiding individuals whereas the likelihood of being sued is uniform across the population. Thus, libel damages serve to transfer resources from agents engaging in misconduct (whose welfare weight is smaller) towards agents who abide by the norms (whose welfare weight is higher), thereby enhancing welfare. By assuming that $\beta = 1$, due to the lack of a re-distributive motive in the welfare specification, the analysis focuses on the behavioral (Pigouvian) efficiency-enhancing impact of libel damages on shaming and the level of misconduct. The social planner is hence solving the following maximization program:

(23) $max_d W(d) \equiv$

$$\int_0^{\hat{\theta}(d)}[y+b-\theta-q\cdot h(d)\cdot S]d\theta+\left[1-\hat{\theta}(d)\right]\cdot[y-(1-q)\cdot h(d)\cdot S]-\frac{\alpha}{2}\cdot\left[\hat{\theta}(d)\right]^2,$$

[7] If, as discussed in footnote 6, the court can verify, with probability 1, whether the plaintiff has been engaged in misconduct, conditions (17) and (22) will be, respectively, replaced by:

$\hat{\theta} = b-(2q-1)\cdot h\cdot S-(1-q)\cdot h\cdot d,$

$h \equiv 1-\frac{(1-q)\cdot\left(1-\hat{\theta}\right)}{(1-q)\cdot\left(1-\hat{\theta}\right)+q\cdot\hat{\theta}}\cdot(F+d)$

Thus, libel damages serve to reduce the extent of shaming, by effectively increasing the cost of reputation associated with false shaming. The effect of libel damages on the level of misconduct is however mixed. Libel damages directly contribute to the reduction in misconduct, as only those who abide by the norms file libel suits and win with certainty. However, they also contribute indirectly to the increase in misconduct, due to the induced reduction in the extent of shaming.

where $\hat{\theta}(d)$ and $h(d)$ denote the solution for the system of equations (17) and (22) as a function of d.[8] We examine whether starting from the *laissez-faire* allocation absent of intervention (d=0) introducing the possibility to file libel suits (d>0 and small) may serve to enhance welfare. Assuming for simplicity that $\hat{\theta}(0) < 1/2$, it follows that for small values of d>0 individuals who engage in misconduct would not file libel suits.[9] It hence follows that (17) and (22) can be reformulated to obtain:

(17') $\quad \hat{\theta} = b - (2q-1) \cdot h \cdot S,$

and

(22') $\quad h \equiv 1 - \frac{(1-q)\cdot(1-\hat{\theta})}{(1-q)\cdot(1-\hat{\theta})+q\cdot\hat{\theta}} \cdot F - \frac{(1-q)/4}{\hat{\theta}\cdot q+(1-\hat{\theta})\cdot(1-q)} \cdot d,$

Fully differentiating the system of two equations, (17') and (22'), with respect to d, yields that $\frac{\partial h}{\partial d} < 0$ and $\frac{\partial \hat{\theta}}{\partial d} > 0$. Thus, introducing libel suits would, as anticipated, serve to reduce the extent of shaming but would consequently induce an increase in the level of misconduct. By virtue of (17') and (22'), one can reformulate the maximization problem given in (23) as follows:

(23') $\quad max_d W(d) \equiv$

$$\int_0^{\hat{\theta}[h(d)]} [y + b - \theta - q \cdot h(d) \cdot S] d\theta + \left[1 - \hat{\theta}[h(d)]\right] \cdot [y - (1-q) \cdot h(d) \cdot S] - \frac{\alpha}{2} \cdot \left[\hat{\theta}[h(d)]\right]^2,$$

where $\hat{\theta}[h(d)] = b - (2q-1) \cdot h(d) \cdot S$.

As misconduct is not directly affected by the introduction of libel suits, but only indirectly via the latter's impact on the extent of shaming in equilibrium, we can assess the desirability of libel suits by examining whether in the benchmark allocation, shaming is over- or under-provided from the social planner's perspective. In case the extent of shaming in the *laissez-faire* allocation is excessive, regulation via setting positive damages for libel suits would be socially desirable.

[8] As $\beta = 1$, the level of libel damages, d, only appears in the welfare function via its behavioral impact on shaming and the level of misconduct.

[9] A sufficient condition for the level of misconduct to be bounded below ½ is that b<1/2.

Differentiating $W(h)$ given in (12') with respect to h, evaluating the derivative at the 'laissez-faire' level of shaming, $h(0)$, yields upon re-arrangement:

(24) $$\left.\frac{\partial W}{\partial h}\right|_{h=h(0)} =$$

$$-\left[\hat{\theta}(h(0)) \cdot q \cdot S + \left(1-\hat{\theta}(h(0))\right) \cdot (1-q) \cdot S\right] + \left[\alpha \cdot \hat{\theta}(h(0)) \cdot (2q-1) \cdot S\right]$$

Introducing libel damages would be socially desirable when the derivative in (24) is negatively signed. Following some algebraic manipulations, one can show that:

(25) $$\left.\frac{\partial W}{\partial h}\right|_{h=h(0)} < 0 \leftrightarrow \hat{\theta}(h(0)) \cdot (\alpha-1) \cdot (2q-1) < (1-q)$$

As $\hat{\theta}(h(0)) > 0$ and 1/2<q<1, it immediately follows that the second inequality condition in (25) holds when $\alpha \leq 1$. Assuming, instead, that $\alpha > 1$, recalling that by presumption $\hat{\theta}(h(0)) < 1/2$, to prove that the second inequality condition in (25) holds it suffices to show that:

(26) $$\frac{1}{2} \cdot (\alpha-1) \cdot (2q-1) < (1-q) \leftrightarrow \alpha < 1/(2q-1),$$

where $\frac{1}{2q-1} > 1$ as 1/2<q<1.

We thus conclude that $\left.\frac{\partial W}{\partial h}\right|_{h=h(0)} < 0$ when $\alpha < 1/(2q-1)$. It is easy to observe that for sufficiently small values of α and q (which is bounded from below by ½) $\left.\frac{\partial W}{\partial h}\right|_{h=h(0)} < 0$. The interpretation is straightforward. Introducing libel suits serves to reduce the extent of shaming but entails the cost of increased misconduct. When either the observed signal is fuzzy and hence not informative (q is small and approaches ½), rendering the deterrence/enforcement effect of enhanced shaming quite limited, or the social cost of misconduct is small (α is small), rendering the social gain from misconduct reduction via increased level of shaming less pronounced, reducing the extent of shaming is socially desirable as it contributes to mitigating the social stigmatizing costs of shaming but bears a relatively small impact on the social cost of misconduct.

It is also straightforward to observe from the second inequality condition in (25) that for sufficiently high values of α and q, recalling that q is bounded from above by 1, $\left.\frac{\partial W}{\partial h}\right|_{h=h(0)} > 0$. Thus, when either the deterrence effect of enhanced shaming is pronounced (the signal is very

informative), or the social cost of misconduct is large, hence the enforcement-enhancing effect of shaming contributes significantly to the reduction in the social cost of misconduct, introducing libel suits which serves to reduce the extent of shaming (and thereby to increase the level of misconduct) is socially undesirable.

5 Truth as a Complete Defense against Defamation

A common feature of defamation law is the view that defendants, in a civil action for defamation, are not liable for damages if they can prove the truth of their apparently defamatory statements. We have thus far assumed, accordingly, that libel damages are exclusively awarded to an individual, who has been falsely subject to shaming by his community member (on the grounds of violating some social norms). That is, the court was ruling that the individual has not been engaged in the act of misconduct alluded to in the defamatory statements. A truthful defamatory statement, in contrast, does not confer any damages to the plaintiff.

Our analysis alludes to the fact that viewing truth as a complete defense against defamation may be socially undesirable, challenging the conventional wisdom and common practice in civil law. To illustrate the point in its sharpest relief, we invoke some strong assumptions about the extent to which both the agents who engage in shaming and the court are informed about the act of misconduct in case. We consider a scenario that, *prima facie*, strongly supports the prevailing doctrine. We assume that the signal observed by the member of the community (based on which he decides whether to 'shame') is fully informative, i.e., q=1. We further assume that the court can verify with probability 1, whether an agent has been engaged in misconduct. Combining the two assumptions implies that only offenders of social norms are subject to shaming, and further implies that libel damages are not awarded, and hence libel suits are not being filed.

Substituting q=1 in the formulae given in footnote 7 implies that in the *laissez-faire* equilibrium:

$\hat{\theta} = b - S,$

$h \equiv 1$

As there are no reputational costs entailed by shaming (the information on which shaming is based is perfectly accurate, by assumption) the extent of shaming in the market equilibrium is the maximal feasible (i.e., universal shaming)

Now assume that $\alpha < \beta$; that is, the social gain from reducing misconduct is smaller than the stigma costs borne by individuals who engage in it. As shown in Section 3, under this parametric assumption, the socially optimal level of shaming that would maximize social welfare would be given by $h = 0$, a corner solution with no shaming. When the negative externalities associated with shaming exceed the positive externalities arising from deterrence, therefore, the market delivers an excessive level of shaming in equilibrium. In this case, awarding damages to offenders who have been subjected to shaming can enhance social welfare by operating as a Pigouvian tax that internalizes the net negative externality generated by shaming. Indeed, the assumption that $\alpha < \beta$, which sets an upper bound on the social gain from shaming, finds implicit support in certain legal systems outside the US. In most European jurisdictions, for instance, truth is not an absolute defense to defamation or reputation-related claims. Courts instead apply a public-interest or necessity qualification even where the facts are true, balancing the value of disclosure against the resulting harm to reputation or privacy.[10]

The prevailing doctrine of restricting libel damages to cases with false shaming is implicitly invoking a strong and morally disconcerting assumption that β, the social weight assigned to offenders (those engaged in misconduct) in the welfare calculus, is universally and categorically lower than α, the social gain from shaming, irrespective of the severity of the act of misconduct (reflected in the value of α). As we argue below, the latter stands in sharp contrast to the conventional approach in law and economics and the common practice.

To see this formally, assume a perfectly informative signal (q=1) and recall that with perfectly accurate information, the market equilibrium yields maximal (i.e., universal) shaming, given by h=1. As shaming is perfectly targeted towards those individuals who engaged in misconduct, they are the only ones to suffer disutility from the entailed stigmatization. Hinging on the principle of

[10] The European Court of Human Rights has repeatedly held that truth does not automatically justify publication. See Von Hannover v. Germany (No. 1), 40 Eur. H.R. Rep. 1 (2004), and Von Hannover v. Germany (No. 2), 55 Eur. H.R. Rep. 15 (2012) (holding that publication of accurate photographs of Princess Caroline's private activities violated Art. 8 because they contributed nothing to a debate of public interest); M.L. & W.W. v. Germany, App. Nos. 60798/10 & 65599/10, 66 Eur. H.R. Rep. 16 (2018) (holding that continued online access to true reports of past crimes violated applicants' right to reintegration and private life); and Axel Springer AG v. Germany, App. No. 39954/08, 55 Eur. H.R. Rep. 6 (2012) (protecting publication of truthful information about a celebrity's arrest because it contributed to a matter of public concern).

truth as a complete defense that renders truthful shaming utterly immune from damages, suppose that libel damages are universally excluded regardless of the severity of the act of misconduct. For this to be the welfare maximizing outcome it necessarily follows that $\beta \leq \alpha$ for every positive α. We assume that the social welfare weight assigned to offenders is represented by a continuously differentiable function of the severity of the act of misconduct, $\beta(\alpha)$. We further assume, plausibly, that the function is non-increasing, that is: $\frac{\partial \beta(\alpha)}{\partial \alpha} \leq 0$.[11] Thus, committing an act of misconduct generally entails some 'social sanction' reflected in being assigned with a social welfare weight smaller than that assigned to norm-abiding individuals ($\beta < 1$). This 'social sanction' is (weakly) increasing in the severity of the act of misconduct, reflecting a corresponding higher extent of social resentment and resignation from the act by the community members. As, by presumption, $\beta \leq \alpha$ for every positive α, it follows that $\beta \to 0$ as $\alpha \to 0$, otherwise, by continuity, $\beta > \alpha$ for some sufficiently low (but strictly positive) value of α, which would contradict our presumption. However, as $\frac{\partial \beta(\alpha)}{\partial \alpha} \leq 0$ it follows that $\beta(\alpha) = 0$ for all levels of α. Thus, a zero welfare-weight is assigned categorically to offenders in the social welfare function. Assigning, categorically, a zero-welfare weight to offenders, however, stands in sharp conflict with Gary Becker's seminal contribution to the literature in law and economics. Becker (1968) suggests that offenders should be assigned a **positive weight** in the social welfare function. Becker (1968) assumes that the government is seeking to set the optimal degree of enforcement (which determines the number of offences in equilibrium) by minimizing a loss function composed of three key components: (i) the damage caused by offences, measured by the difference between the amount of harm caused to the members of society other than the offenders minus the **social value** of the **gain to offenders**; (ii) the social cost of apprehension and conviction; and, (iii) the social cost of punishment. Notably, Becker's framework explicitly includes the social value of gains to offenders within the welfare calculus. The term "social value" suggests the possibility that the gains from offending may be weighted differently than offenders themselves would assign them, potentially allowing for a discounting of their welfare in the broader social welfare function. Nevertheless, a positive social value of the gain to offenders is included in the welfare calculus,

[11] By assuming that the function is non-increasing we accommodate, as a special case, the possibility that the social welfare weight is set at some fixed level and hence is independent of the severity of the act of misconduct.

namely, $\beta > 0$. Becker makes this argument in relation to criminal offenses. A fortiori, this applies to our case, which involves the violation of social norms that does not rise to the level of a crime.

Invoking some additional assumptions with respect to the properties of the function $\beta(\alpha)$, it is straightforward to derive a simple modified policy rule that, in the case where libel suits concern truthful defamatory statements, is welfare-enhancing relative to the prevailing doctrine. Formally, we assume the following: (i) $\beta \to 1$ as $\alpha \to 0$; and (ii) $\beta \to 0$ as $\alpha \to \overline{\alpha}$, with $\overline{\alpha} > 0$ denoting the maximal level of severity associated with the act of misconduct (defining an upper bound on the social gains from shaming). Letting $\Delta(\alpha) \equiv \alpha - \beta(\alpha)$, it follows immediately that $\frac{\partial \Delta(\alpha)}{\partial \alpha} > 0$. Moreover, $\Delta \to -1$ as $\alpha \to 0$ and $\Delta \to \overline{\alpha}$ as $\alpha \to \overline{\alpha}$. By the continuity of $\Delta(\alpha)$, employing the intermediate value theorem, it follows that there exists some level of α, $0 < \hat{\alpha} < \overline{\alpha}$, such that $\Delta(\hat{\alpha}) = 0$. Finally, as $\frac{\partial \Delta(\alpha)}{\partial \alpha} > 0$, $\hat{\alpha}$ is unique.

We thus conclude that $\Delta(\alpha) > 0$ if-and-only-if $\alpha > \hat{\alpha}$. The 'complete' defense of truth against defamation should be hence restricted to acts of misconduct that are deemed sufficiently severe from the perspective of the society. The latter (normative) threshold is likely to vary across societies and reflect the prevailing norms in the community, the personal characteristics of the offenders and the context in which the act of misconduct that has been committed.

6.1 Heterogenous Signals[12]

Suppose that members of the community differ in the amount of information they possess. For simplicity, assume that for each agent *i*, a fraction δ_t of *i*'s community members observes a signal *t* that obtains, with the probability $\frac{1}{2} < q_t < 1$, a high realization if *i* commits an act of misconduct and a low realization if *i* abides by the norm. Suppose further that there are two signals, *t*=1,2, so that $\delta_1 + \delta_2 = 1$ and, for simplicity, that the type of signal is independently and identically

[12] In this subsection we consider heterogeneity in the quality of information possessed by community members. One may also consider other sources of heterogeneity. One notable example is heterogeneity in the severity of the acts of misconduct committed by the members of the community. Other things being equal, the socially desirable level of shaming should correspond with the severity of the misconduct committed. The social planner may, accordingly, set a differential system of damages in libel suits based on the category of misconduct (e.g., distinguishing between minor acts of misconduct, such as littering in the public domain, and major acts of misconduct, such as sexual harassment). This point relates to our discussion above of the issue of non-proportionality associated with shaming, namely the potential weak correlation between private incentives to engage in shaming and the severity of the act of misconduct.

distributed across the members of the community. The latter assumption implies, plausibly, that members of the community may be more informed with respect to some agents and less informed with respect to others. Finally, assume that $q_2 > q_1$. Thus, the signal *t*=2 is more informative than *t*=1. The heterogeneity in the quality of the signal may be driven by variation in the access to information and/or validation opportunities. Based on the realization of the observed signal (associated with each agent *i*), each community member decides whether to engage in shaming of agent *i*. By engaging in shaming each such member exerts two forms of externalities on the community: (i) a positive externality due to the induced deterrence effect and the corresponding reduction in the level of misconduct; (ii) a negative externality due to the stigmatization of other members of the community. As discussed in section 3 above, the more accurate the signal becomes, the larger the social net benefit from enhanced shaming (accounting for both conflicting external effects) turns out to be. To internalize the combined external effect exerted by each member of the community who engages in shaming, one would ideally use a system of differential *Pigouvian* 'taxes' and 'subsidies' on shaming, setting a distinct tax/subsidy for each type of signal. However, signal types are plausibly assumed to be private information unobservable by the government. In the presence of asymmetric information, hence, the government is typically unable to implement the First-Best solution and must compromise on the Second-Best optimum (setting a universal *Pigouvian* tax/subsidy rather than a system of type-dependent taxes and subsidies, as it would ideally aim to do). It turns out, however, that relying on uniform libel damages would still enable the government to implement the First-Best optimum. We turn next to illustrate this point.

Consider for concreteness a scenario where $q_2 = 1 - \epsilon$ and $q_1 = \frac{1}{2} + \epsilon$, with $\epsilon > 0$ and small. That is, an individual observing the realization of the signal *t*=2 associated with some agent *i* is highly informed with respect to *i*. In contrast, an individual observing the realization of the signal *t*=1 associated with *i* is virtually uninformed with respect to *i*. Let h_t^* , *t*=1,2, denote the extent of shaming generated by members of the community that observe the signal *t*, and $\hat{\theta}$ denotes the level of misconduct, under the 'laissez-faire' (unregulated) market equilibrium. Following our analysis in section 2, the equilibrium would be given by the solution to the following system of three equations for three unknowns (h_1^*, h_2^* and $\hat{\theta}$):

(27) $\quad \hat{\theta} = b - [\delta_1 \cdot (2q_1 - 1) \cdot h_1^* + \delta_2 \cdot (2q_2 - 1) \cdot h_2^*] \cdot S$

(28) $h_1^* = 1 - \frac{(1-q_1)\cdot(1-\widehat{\theta})}{(1-q_1)\cdot(1-\widehat{\theta})+q_1\cdot\widehat{\theta}} \cdot F$

(29) $h_2^* = 1 - \frac{(1-q_2)\cdot(1-\widehat{\theta})}{(1-q_2)\cdot(1-\widehat{\theta})+q_2\cdot\widehat{\theta}} \cdot F$

As $\frac{1}{2} < q_1 < q_2 < 1$ and $0 < b, F < 1$, it follows from conditions (27)-(29) that in equilibrium $0 < h_1^* < h_2^* < 1$. Thus, the extent of shaming generated by an 'informed' member of the community strictly exceeds that generated by an 'uninformed' one.

Modifying the definition of the social welfare function for the case with two types of signals, maintaining our simplifying assumption that $\beta = 1$, yields:

(30) $W(h_1, h_2) \equiv \int_0^{\widehat{\theta}(h_1,h_2)} [y + b - \theta - (\delta_1 \cdot q_1 \cdot h_1 + \delta_2 \cdot q_2 \cdot h_2) \cdot S] d\theta$

$$+\left[1 - \hat{\theta}(h)\right] \cdot [y - [\delta_1 \cdot (1 - q_1) \cdot h_1 + \delta_2 \cdot (1 - q_2) \cdot h_2] \cdot S] - \frac{\alpha}{2} \cdot [\hat{\theta}(h)]^2$$

Assuming that the social cost of misconduct is significant ($\alpha > 1$), following our analysis in section 3, the socially desirable extent of shaming generated by an 'informed' member of the community and an 'uninformed' one, respectively, obtained by the maximization of social welfare function in (30), would be given by $h_1 = 0$ ('no shaming') and $h_2 = 1$ ('universal shaming'). Thus, as $h_2^* < h_2$, 'informed' agents should be subsidized to enhance the extent of shaming due to the (combined) positive externality they confer (gains from reduction in misconduct outweigh the cost of stigmatization), whereas, as $h_1^* > h_1$, 'uninformed' agents should be taxed due to the (combined) negative externality they exert (the cost of stigmatization exceeds the gains from enhanced deterrence), serving to reduce the extent of shaming.

More generally, as $q_2 > q_1$, 'informed' agents should be typically subject to a lower tax rate (or conferred with a more generous subsidy) relative to their 'uninformed' counterparts. To implement a differential '*Pigouvian'* system, assuming the signal types are private information, one can use a combination of libel-damages and a direct subsidy/tax of shaming.[13] Notice that the likelihood of being subject to a libel suit is diminishing in q [which follows from (22')]. Thus, the

[13] We assume that the costs of the subsidy are financed by a lump sum tax levied across the board on all members of the community. If a direct tax is being levied on shaming, the tax revenues are assumed to be rebated in a lump-sum fashion to the community members.

combination of the two instruments can indeed implement the socially desirable differential *Pi*gouvian system. To see this, notice that the effective tax levied on 'informed' agents would be smaller although damages are universal due to the difference in the likelihood of being subject to a libel suit. This latter feature would allow the government to implement the socially desirable level of differentiation in the incentives given to 'informed' and 'uninformed' agents. The direct subsidy/tax would serve as a level shifter to calibrate to the socially desirable aggregate extent of shaming.

Formally, denoting by *d* the universal level of libel damages and by *s* the universal level of a direct shaming subsidy, and letting h_1 and h_2 denote the socially desirable extent of shaming associated with an 'uninformed' agent and an 'informed' one, respectively, the First Best optimum is given by the solution to the following system of three equations:

(31) $\hat{\theta} = b - [\delta_1 \cdot (2q_1 - 1) \cdot h_1 + \delta_2 \cdot (2q_2 - 1) \cdot h_2] \cdot S$

(32) $h_1 = 1 - \frac{(1-q_1)\cdot(1-\hat{\theta})}{(1-q_1)\cdot(1-\hat{\theta})+q_1\cdot\hat{\theta}} \cdot F + s - \frac{(1-q_1)/4}{\hat{\theta}\cdot q_1+(1-\hat{\theta})\cdot(1-q_1)} \cdot d$

(33) $h_2 = 1 - \frac{(1-q_2)\cdot(1-\hat{\theta})}{(1-q_2)\cdot(1-\hat{\theta})+q_2\cdot\hat{\theta}} \cdot F + s - \frac{(1-q_2)/4}{\hat{\theta}\cdot q_2+(1-\hat{\theta})\cdot(1-q_2)} \cdot d.$

Notice that by assumption $\beta = 1$ and the government budget is fiscally balanced by a lump-sum tax/transfer, hence the level of libel damages, *d*, and the rate of shaming subsidy (tax, if negative), *s*, play no re-distributive role, and only serve to implement the socially desirable levels of shaming given by h_1 and h_2.

Notice further that the solution for the system (31)-(33) is feasible only when $d \geq 0$ (libel damages are non-negative). Subtracting (32) from (33) yields a sufficient condition for feasibility, given by:

(34) $h_2 - h_1 \geq \left(\frac{(1-q_1)\cdot(1-\hat{\theta})}{(1-q_1)\cdot(1-\hat{\theta})+q_1\cdot\hat{\theta}} - \frac{(1-q_2)\cdot(1-\hat{\theta})}{(1-q_2)\cdot(1-\hat{\theta})+q_2\cdot\hat{\theta}}\right) \cdot F,$

which holds for $F>0$ sufficiently small.

Notice that *s* is unrestricted and could be either positive or negative (in which case it forms a tax). In the latter case the combined net externality generated by both 'informed' and 'uninformed' shaming is negative; namely, the negative externality associated with stigmatization exceeds the positive externality associated with enhanced deterrence of social misconduct.

In our parametric example, for instance, $h_2 - h_1 = 1$, hence, as by presumption, as $0<F<1$, it is straightforward to verify that the condition in (34) is satisfied (as a strict inequality). Furthermore, $s>0$, as the combined net externality associated with 'informed' agents is positive. The First Best is hence attainable, despite the asymmetric information between the agents and the government.

The presence of reputational costs serves to some extent to induce stronger disincentives for the 'un-informed' agents to engage in shaming. However, if the reputational incentives are sufficiently moderate, the net positive externality generated by 'informed' shaming is larger than that generated by 'uninformed' shaming (or the net negative externality is smaller), which implies that the combination of a shaming subsidy/tax and libel damages can do the job and implement the socially desirable outcome.

Notice that in case there are more than two types of signals, the two linear instruments would not suffice to implement the optimal differential *Pigouvian system* (setting a distinct tax/subsidy for each signal type). In such a case, a non-linear incentive compatible scheme would be required to attain the second best optimum.[14]

6.2 Herding and Indeterminacy

The act of shaming may reflect other-regarding preferences and prevailing social norms. The likelihood of being engaged in shaming may plausibly increase when shaming is more prevalent in the community and decrease when it becomes less common. Such herding patterns can be readily embedded in our basic setup, as will be shown in what follows.

Assume that the benefit from posting a 'dislike' is given by:

(35) $\quad B = K \cdot (1 - \sigma) + \sigma \cdot h$

Thus, the benefit from engaging in shaming is given by a weighted average of the 'intrinsic' benefit from exposure, K, and the herding component, h (the extent of shaming). The relative weight

[14] The government would offer a menu of (s, d) pairs, where s denotes a direct subsidy to shaming and d denotes the level of libel damages, (s_t, d_t) t=1,2…T, with a higher t reflecting a more informative signal, such that $s_t<s_{t'}$ and $d_t<d_{t'}$ for $t'>t$, from which individuals will self-select. Naturally, the menu would be designed in a manner that would induce higher types (possessing better information) to choose pairs offering a higher subsidy rate combined with higher libel damages (the menu would satisfy incentive compatibility conditions).

assigned to the herding component, $0 < \sigma < 1$, reflects the magnitude of the herding pattern. Notice, that in the absence of herding, that is when $\sigma = 0$, the model reverts to the baseline setup, in which *B*=*K*. The formulation in (35) captures, in a simple form, the fact that the individual's incentives to engage in shaming become stronger as the extent of shaming in the community increases. Under the new formulation, agents engage in shaming if-and-only-if:

(36) $K \cdot (1-\sigma) + \sigma \cdot h > \hat{p}[\hat{\theta}(h)] \cdot F,$

where $\hat{p}(\hat{\theta}) \equiv \frac{(1-q)\cdot(1-\hat{\theta})}{(1-q)\cdot(1-\hat{\theta})+q\cdot\hat{\theta}}$ and $\hat{\theta}(h) = b - (2q-1) \cdot h \cdot S$.

Suppose further that the extent of herding is sufficiently pronounced, so that σ is sufficiently large and satisfies:

(37) $\sigma > \max\left(\hat{p}\left[\hat{\theta}(1)\right] \cdot F, 1 - \hat{p}[\hat{\theta}(0)] \cdot F\right).$

One can show that there are two locally stable equilibria which are given by the two corner solutions: 'no shaming' ($h^* = 0$) and 'universal shaming' ($h^* = 1$). There is also an additional (locally unstable) unique interior equilibrium, given by some $h^* \in (0,1)$.

To see this, notice that by virtue of the inequality condition in (36), an equilibrium is defined by a cutoff, $0 \leq K^* \leq 1$, such that agents engage in shaming if-and-only-if $K > K^*$. As by assumption *K* is drawn independently from a uniform distribution with support [0,1], in equilibrium, it follows that $K^* = 1 - h^*$, where h^* denotes the extent of shaming in equilibrium.

In a 'no-shaming' equilibrium, with h^*=0, the cutoff is given by $K^* = 1$, and the inequality condition in (36) satisfies:

(38) $K^* \cdot (1-\sigma) + \sigma \cdot h^* < \hat{p}\left[\hat{\theta}(h^*)\right] \cdot F \leftrightarrow \sigma > 1 - \hat{p}\left[\hat{\theta}(0)\right] \cdot F,$

where the second inequality is obtained by substituting for $h^* = 0$ and $K^* = 1$, and re-arranging. The inequality in (38), which follows from condition (37), guarantees that none of the agents engages in shaming, consistent with the presumption that $h^* = 0$.

In a 'universal shaming' equilibrium, with h^*=1, the cutoff is given by $K^* = 0$, and the inequality condition in (36) satisfies:

(39) $K^* \cdot (1-\sigma) + \sigma \cdot h^* > \hat{p}[\hat{\theta}(h^*)] \cdot F \leftrightarrow \sigma > \hat{p}[\hat{\theta}(1)] \cdot F,$

where the second inequality follows by substituting for $h^* = 1$ and $K^* = 0$, and rearranging. The inequality in (39), which follows from condition (37), guarantees that all agents engage in shaming, consistent with the presumption that $h^* = 1$.

Finally, in an interior solution, $0 < h^* < 1$, and the cutoff $K^* = 1 - h^*$ satisfies the condition in (36) as an equality, which yields upon re-arrangement:

(40) $\quad (1-\sigma) + (2\sigma - 1) \cdot h^* = \hat{p}[\hat{\theta}(h^*)] \cdot F.$

The equality condition in (40) guarantees that a fraction of the agents engages in shaming (so that the marginal agent is indifferent between engaging in shaming and refraining from doing so), consistent with the presumption that $0 < h^* < 1$.

We turn next to show that the interior equilibrium is well defined and unique.

Let $G(h) \equiv (1-\sigma) + (2\sigma - 1) \cdot h - \hat{p}[\hat{\theta}(h)] \cdot F$, where $G(h)$, defined over the interval [0,1], measures the net benefit from engaging in shaming by the marginal agent. By virtue of (38) it follows that $G(0)<0$, whereas, by virtue of (39) it follows that $G(1)>0$. Thus, by the continuity of G, applying the Intermediate Value Theorem, it follows that there exists some $0 < h^* < 1$ for which $G(h^*)=0$. Thus, the interior equilibrium is well defined.

Differentiating G twice, it follows that $G''(h) < 0$. The strict concavity of G implies that G has <u>at most</u> two roots in the interval [0,1]. Now, suppose by negation that G has indeed two roots and denote them respectively by $0 < h' < h'' < 1$, so that $G(h') = G(h'') = 0$ and $G(h) \neq 0$ otherwise. By virtue of the strictly concavity of G and as $G(0)<0$ it follows that $G'(h') > 0$ and $G'(h'') < 0$. Thus, $G'(h) < 0$ for all $h > h''$, by the strict concavity of G, implying that $G(h) < 0$ for all $h > h''$. This yields a contradiction to the fact that $G(1)>0$. We thus conclude that G has a single root in the interval [0,1] and hence h^* is unique.

Finally, notice that the (local) stability of the two corner equilibria and the (local) instability of the 'interior equilibrium' follow from the fact that $G(h) < 0$ for all $h < h^*$ and $G(h) > 0$ for all $h > h^*$. The extent of shaming would hence decrease towards $h=0$ when $h < h^*$ and would correspondingly increase towards $h=1$ when $h > h^*$, as, by the definition of G, the marginal agent would prefer to engage in shaming when $G(h) > 0$ and to refrain from doing so when $G(h) < 0$.

Several observations are in order. First notice, that the social planner's problem in the presence of herding patterns is not confined to the internalization of externalities not accounted for by the agents but concerns also the selection of the desirable equilibrium configuration (as there are generically multiple such equilibria). Moreover, notice that in the presence of herding, there is a substantial amplification of the 'warm glow' effect of shaming, which pushes the economy into extreme configurations. The latter amplifies the social cost entailed by under- or over provision of shaming. Finally, and perhaps most importantly, the existence of multiple (locally) stable equilibria results in an undesirable indeterminacy, which implies that the same type of act of misconduct may trigger different outcomes in terms of the level of shaming and the extent of stigma suffered by an individual violating the social norms. For instance, assuming the inequality condition in (37) holds, two identical communities (in terms of the distribution of preferences and information) may end up in utterly different (polarized) equilibrium configurations (no-shaming and universal-shaming). The notion of determinacy lies at the essence of criminal law and the need for public provision of enforcement, and the lack of which is often raised by critiques of shaming as an alternative legitimate market enforcement tool. Ensuring that the economy would consistently coordinate on the same equilibrium is hence essential for the adoption of shaming.

One simple way to address the 'network externalities' associated with herding is to introduce a *Pigouvian* tax which offsets these externalities and thereby eliminates the 'threat' of indeterminacy. To illustrate the point, we consider again the setup with two types of signals in which we embed herding externalities (we maintain the notation from subsection 5.1). Denote by h_1 and h_2 the extent of shaming associated with an 'uninformed' agent and an 'informed' one, respectively. We assume that the benefit from posting a 'dislike' is given by:

(41) $\quad B = K \cdot (1-\sigma) + \sigma \cdot (\delta_1 \cdot h_1 + \delta_2 \cdot h_2),$

where $0 < \sigma < 1$ reflects, as before, the relative weight assigned to the herding component (given now by the average extent of shaming) and *K*, the intrinsic benefit from exposure, is assumed to be drawn from a uniform distribution over the support [0,1]. An agent who observes a high realization of a signal *t*, *t*=1,2, engages in shaming if-and-only-if:

(42) $\quad K \cdot (1-\sigma) + \sigma \cdot (\delta_1 \cdot h_1 + \delta_2 \cdot h_2) > \hat{p}_t[\hat{\theta}(h_1, h_2)] \cdot F,$

where,

$\hat{p}_t(\hat{\theta}) \equiv \frac{(1-q_t)\cdot(1-\hat{\theta})}{(1-q_t)\cdot(1-\hat{\theta})+q_t\cdot\hat{\theta}}$ and $\hat{\theta}(h_1,h_2) = b - [\delta_1 \cdot (2q_1 - 1) \cdot h_1 + \delta_2 \cdot (2q_2 - 1) \cdot h_2] \cdot S$.

We turn next to demonstrate that the social optimum can be implemented by a combination of a direct tax/subsidy levied on shaming, $s(h_1, h_2)$, and libel damages, *d*. Notice, crucially, that the tax/subsidy is set as a function of the levels of shaming in the community (and is not flat as in the baseline model). This latter feature serves to offset the 'herding externality' and the resulting indeterminacy.

Let $\tilde{h}_1$ and $\tilde{h}_2$ denote the socially desirable extent of shaming associated with an 'uninformed' agent and an 'informed' one, respectively. The social optimum is given by the solution to the following system of three equations (for three unknowns: s, d and $\hat{\theta}$):

(43) $$\hat{\theta} = b - [\delta_1 \cdot (2q_1 - 1) \cdot \tilde{h}_1 + \delta_2 \cdot (2q_2 - 1) \cdot \tilde{h}_2] \cdot S$$

(44) $$\left(1 - \tilde{h}_1\right) \cdot (1 - \sigma) + \sigma \cdot \left(\delta_1 \cdot \tilde{h}_1 + \delta_2 \cdot \tilde{h}_2\right) =$$

$$\frac{(1 - q_1) \cdot \left(1 - \hat{\theta}\right)}{(1 - q_1) \cdot \left(1 - \hat{\theta}\right) + q_1 \cdot \hat{\theta}} \cdot F - s\left(\tilde{h}_1, \tilde{h}_2\right) + \frac{(1 - q_1)/4}{\hat{\theta} \cdot q_1 + \left(1 - \hat{\theta}\right) \cdot (1 - q_1)} \cdot d$$

(45) $$\left(1 - \tilde{h}_2\right) \cdot (1 - \sigma) + \sigma \cdot \left(\delta_1 \cdot \tilde{h}_1 + \delta_2 \cdot \tilde{h}_2\right) =$$

$$\frac{(1 - q_2) \cdot \left(1 - \hat{\theta}\right)}{(1 - q_2) \cdot \left(1 - \hat{\theta}\right) + q_2 \cdot \hat{\theta}} \cdot F - s\left(\tilde{h}_1, \tilde{h}_2\right) + \frac{(1 - q_2)/4}{\hat{\theta} \cdot q_2 + \left(1 - \hat{\theta}\right) \cdot (1 - q_2)} \cdot d$$

where $s\left(\tilde{h}_1, \tilde{h}_2\right) = s - \sigma \cdot \left(\delta_1 \cdot \tilde{h}_1 + \delta_2 \cdot \tilde{h}_2\right)$.

Following our argument from subsection 5.1, the solution for the system (43)-(45) is feasible only when $d \geq 0$. A sufficient condition for this is:

(46) $$\tilde{h}_2 - \tilde{h}_1 \geq \frac{1}{(1-\sigma)} \cdot \left(\frac{(1-q_1)\cdot(1-\hat{\theta})}{(1-q_1)\cdot(1-\hat{\theta})+q_1\cdot\hat{\theta}} - \frac{(1-q_2)\cdot(1-\hat{\theta})}{(1-q_2)\cdot(1-\hat{\theta})+q_2\cdot\hat{\theta}}\right) \cdot F,$$

which holds for $F>0$ sufficiently small.

To illustrate the implementation of the social optimum, consider again the parametric example from subsection 5.1, where $q_2 = 1 - \epsilon$ and $q_1 = \frac{1}{2} + \epsilon$, with $\epsilon > 0$ and small. The socially

desirable levels of shaming are given by $\tilde{h}_2 = 1$ and $\tilde{h}_1 = 0$, hence, $\tilde{h}_2 - \tilde{h}_1 = 1$. Assuming further that $0 < F < 1 - \sigma$, it is straightforward to verify that the condition in (46) is satisfied (as a strict inequality). Furthermore, it is easy to verify that $s>0$, by substituting for $\tilde{h}_2 = 1$ into (45).

The possibility to implement the First-Best solution in the presence of asymmetric information hinges on the fact that the government can rely on a type-independent fiscal instrument to internalize the herding externality (hence no incentive compatibility issues arise). The reason for this is the 'atmospheric' nature of the herding externality – the herding component in the payoff function is identical across all agents. If, alternatively, 'informed' agents would care more about shaming patterns amongst their 'informed' counterparts and likewise for 'uninformed' agents, then type-dependent differential instruments would be needed. Incentive compatibility issues would then typically arise and dictate compromising on a Second-Best solution.

7. Conclusion

We have examined a scenario in which each community member faces a dual choice. The first decision revolves around whether to defy social norms, weighing the personal gains from engaging in such behavior against the potential costs of being exposed to stigma through the social network's internal shaming mechanisms. The second choice involves deciding whether to participate in shaming, considering the personal benefits of revealing social misconduct and the reputational costs incurred if the disseminated information is proven false. In the absence of government intervention, we characterize the equilibrium that concurrently determines the level of misconduct and the extent of shaming.

Within this equilibrium, two conflicting externalities arise: (i) a positive externality from the deterrence of misconduct induced by shaming and (ii) a negative externality from the stigma costs borne by individuals subjected to shaming. The presence of these externalities typically leads to a suboptimal level of shaming under a laissez-faire regime (with no government intervention in place).

To explore the normative implications, we define the socially optimal level of shaming as a function of key model parameters, including: the quality of information available to the public for making shaming decisions, the severity of misconduct in terms of its social cost, the relative social

weight assigned to the well-being of shamed violators versus wrongfully shamed (norm-abiding) individuals, and the extent of herding behavior in shaming dynamics.

We then examine potential government interventions to internalize these externalities and mitigate the resulting inefficiencies. In particular, we focus on the role of libel suits and the direct subsidization or taxation of shaming as regulatory tools. We demonstrate that the First-Best solution can be achieved even in the presence of asymmetric information and herding externalities. When herding dynamics are present, multiple equilibria emerge and libel damages or targeted subsidies can restore determinacy.

Most importantly, our analysis challenges the entrenched U.S. principle that truth serves as a complete defense to defamation. We demonstrate that this rule is not welfare-maximizing and may produce excessive social costs in the digital environment, thereby warranting a re-evaluation of defamation law through a public-economics lens.

References

Andreoni, James. 1987. "An Experimental Test of the Public-Goods Crowding-Out Hypothesis." *American Economic Review* 77(5): 891–904.

———. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *Economic Journal* 100(401): 464–477.

Arbel, Yonathan A. 2023. "The Credibility Effect: Defamation Law and Audiences." *Journal of Legal Studies* 52(2): 417–443.

Arbel, Yonathan A., and Murat C. Mungan. 2019. "The Case against Expanding Defamation Law." *Alabama Law Review* 71(2): 453–497.

———. 2023. "Defamation with Bayesian Audiences." *Journal of Legal Studies* 52(2): 445–483.

Axel Springer AG v. Germany. 2012. Application No. 39954/08, 55 European Human Rights Reports 6 (Eur. Ct. H.R.).

Bar-Gill, Oren, and Assaf Hamdani. 2002. "Optimal Liability for Libel." *Harvard John M. Olin Discussion Paper Series*, Discussion Paper No. 372, Harvard Law School, Cambridge, MA.

———. 2003. "Optimal Liability for Libel." *Contributions to Economic Analysis & Policy* 2(1): 1–26.

Charter of Fundamental Rights of the European Union. 2012. Official Journal of the European Union C 326/391, 26 October 2012.

Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (Convention 108+). 2018. Council of Europe Treaty Series No. 223 (Modernised Convention 108).

Cooter, Robert. 2000. *The Strategic Constitution.* Princeton, NJ: Princeton University Press.

Dalvi, Manoj, and James F. Refalo. 2008. "An Economic Analysis of Libel Law." *Eastern Economic Journal* 34(1): 74–94.

European Convention on Human Rights. 1950. Rome, 4 November 1950, as amended by Protocols Nos. 11 and 14, Council of Europe.

Farber, Daniel A. 1991. "Free Speech without Romance: Public Choice and the First Amendment." *Harvard Law Review* 105(2): 554–583.

Garoupa, Nuno. 1999a. "Dishonesty and Libel Law: The Economics of the 'Chilling' Effect." *Journal of Institutional and Theoretical Economics* 155(2): 284–300.

———. 1999b. "The Economics of Political Dishonesty and Defamation." *International Review of Law and Economics* 19(2): 167–180.

General Data Protection Regulation (GDPR) (Regulation [EU] 2016/679). 2016. Official Journal of the European Union L 119/1, 4 May 2016.

Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González. 2014. Case C-131/12, European Court of Justice (Grand Chamber), ECLI:EU:C:2014:317.

Hemel, Daniel, and Ariel Porat. 2019. "Free Speech and Cheap Talk." *Journal of Legal Analysis* 11(1): 46–103.

Hylton, Keith N. 1996. "A Missing Markets Theory of Tort Law." *Northwestern University Law Review* 90(3): 977–1030.

Klonick, Kate. 2016. "Re-Shaming the Debate: Social Norms, Shame, and Regulation in an Internet Age." *Maryland Law Review* 75(4): 1029–1069.

M.L. and W.W. v. Germany. 2018. Applications Nos. 60798/10 and 65599/10, 66 European Human Rights Reports 16 (Eur. Ct. H.R.).

New York Times Co. v. Sullivan, 376 U.S. 254 (1964).

Posner, Eric A. 2000. *Law and Social Norms.* Cambridge, MA: Harvard University Press.

Posner, Richard A. 1997. “Social Norms and the Law: An Economic Approach.” *American Economic Review* 87(2): 365–369.

Restatement (Second) of Torts § 581A. 1977. St. Paul, MN: American Law Institute.

Restatement of the Law Third, Torts: Defamation and Privacy. 2025. St. Paul, MN: American Law Institute.

Ronson, Jon. 2015. *So You’ve Been Publicly Shamed.* London: Picador.

Sunstein, Cass R. 2020. “Falsehoods and the First Amendment.” *Harvard Journal of Law & Technology* 33(2): 387–426.

United States v. Alvarez, 567 U.S. 709 (2012) (plurality opinion).

Von Hannover v. Germany (No. 1). 2004. 40 European Human Rights Reports 1 (Eur. Ct. H.R.).

Von Hannover v. Germany (No. 2). 2012. 55 European Human Rights Reports 15 (Eur. Ct. H.R.).