Law, Probability and Risk (2018) **17**, 201–223
Advance Access publication on April 11, 2018

doi:10.1093/lpr/mgy003

Free expression and defamation

Nathan Berg[†]
Department of Economics, University of Otago, PO Box 56, Dunedin, New Zealand

and

Jeong-Yoo Kim[‡]
Department of Economics, Kyung Hee University, Hoegi-dong, Dongdaemun-ku, Seoul 130-701, Korea

[Received on 18 November 2016; revised on 9 March 2017; accepted on 16 March 2018]

An economic rationale for restrictions on free expression is considered. An expresser obtains positive utility from expressing something but it may have damaging effects, which can be measured by its *mean squared error*, on others. If social losses from expression exceed benefits significantly, restricting expression can improve social welfare. We analyse expression of distorted information and the social welfare consequences of laws that restrict speech according to three standards: whether transmitted information contains falsity, whether false statements are deliberate and whether the expresser intentionally applies a biased filter to selectively express private information. We also show that the anti-defamation law adopting the negligence rule can lead to the socially optimal level of expression by adjusting the due care standard appropriately.

Keywords: defamation; libel; slander; speech; freedom of expression; constitution.

1. Introduction

Free expression is one of the rights given special consideration in the US Constitution. Many national constitutions similarly prioritize freedom of speech as a core principle that underlies citizens' well-being. The Universal Declaration of Human Rights (UDHR) adopted by the United Nations General Assembly in 1948 stipulates that freedom of thought and expression are universal human rights.[1] However, it is also true that nearly all countries impose restrictions on free expression, applying various rationales.

In this article, we consider an economic rationale for restricting free expression based on the effects of belief accuracy on social welfare, i.e. the extent to which expression affects the gap between individuals' subjective beliefs and the objective state of the world, under the assumption that social welfare improves when individuals correctly perceive what is objectively true.

An individual chooses to express something (that contains some facts) because doing so improves the expresser's own utility;[2] but transmitting expression to others also has the potential to do good or

[†]Email: nathan.berg@otago.ac.nz

[‡]Corresponding author. Email: jyookim@khu.ac.kr

[1] Article 19 of UDHR states: 'Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.'

[2] The idea is that some agents experience positive utility from the act of influencing others.

harm, which is represented by increases or decreases in the utility of individuals who receive an expressed message. If the externalities include utility losses experienced by a third person through a reputation effect, it belongs to defamation. Since there is obviously no social-welfare rationale for restricting expression that increases others' utility, our focus in this article is defamatory statements, i.e. expression that decreases someone else's utility especially by conveying wrong belief about some facts.[3]

Not everyone will agree on what kind of social welfare function should be used to reflect trade-offs between different agents who express and receive messages. In the case of laws that restrict speech, some would argue that the principle of free speech should itself be included as an input into any social welfare function that reasonably represents a society's wellbeing.[4] The two authors of this article disagree, for instance, on whether deviations from the principle of freedom of speech should be netted out as a social loss beyond the individually experienced costs and benefits (i.e. negative externalities that are prevented by laws that restrict speech). For example, one view that prioritizes individual liberty above negative externalities generated by freedom of speech could be captured by a lexicographic social welfare function that prioritizes any loss of freedom to express over the costs that expression imposes on others (e.g. subjective beliefs that embarrass someone else so intensely that the person commits suicide).

Our analysis will focus mostly on the issue of interpersonal trade-offs using a social cost metric that depends on information distortion which can be measured by the 'mean squared error' of an expression. This view is based on the tenet that wellbeing is enhanced when subjective beliefs about the underlying state of nature are objectively accurate. The social welfare function that we consider assumes that social welfare is an increasing function of the extent to which individuals can discover what is true. We will also show how our specification of social welfare, focused primarily on belief accuracy, can be generalized to address the second issue of whether limitations on liberty should be represented as a distinct category of harm or social cost and then integrated among the informational trade-offs that our social welfare function captures.

We proceed to specify social welfare as a simple Benthamite aggregation of individuals' net payoffs with and without anti-defamation laws, under contrasting assumptions about the structure of information in the environment. Equal weighting of individual net payoffs (i.e. without penalizing legal standards that violate principles which some observers would regard as incommensurably more important than the sum of individual payoffs) serves as a methodologically orthodox point of departure for our social welfare analysis, which provides a novel economic characterization of legal standards used in cases involving laws that restrict expression.

Whenever one individual's expression imposes a large negative externality on someone else, our specification allows for the possibility of social welfare improvements from anti-defamation laws that restrict expression. This possibility captures the view of those who advocate for legally restricting speech based on the straightforward proposition: if one agent's gain from free expression is exceeded by (net) social losses, then restricting freedom of expression would be an improvement in social efficiency.[5] This view appears in the work of some legal scholars such as Sunstein (1993, 2014)

[3] It includes both libel (i.e. written defamatory statements) and slander (i.e. spoken defamatory statements). Our analysis is, therefore, restricted only to expression of facts, not expression of opinions.

[4] For more on this view, see Baker's (1989) comparisons of liberty versus marketplace theories. Marketplace theory corresponds to utilitarian welfare theory. In contrast, liberty theory corresponds to anti-utilitarian theory such as Dworkin (1977).

[5] Mill (1978) suggests that rules of conduct are needed to regulate the actions of individuals in a political community. The limitation on free expression that Mill can be understood to accept would be based on what he describes as 'one very simple

and Radin (1996), who emphasize the potential harms caused by free expression, arguing that speech should therefore be subject to controls (under appropriate conditions).[6]

Under what circumstances might this condition rationalizing the restriction of expression hold? We believe that information distortion is at the centre of such disputes. In this article, we focus on two observed regularities among existing laws that restrict expression. First, many countries restrict free expression only when that expression is based on false information. That is, if it is based on a truth or if the expresser believes that her expression is true, then such expressions are not subject to extant anti-defamation law.[7] We analyse the issues of whether expression based on false information should be exempted from restrictions on defamation and whether the intention of an expresser who may have believed that an expressed falsehood was true should be exempt from anti-defamation law.[8]

A second issue that our model addresses concerns the observation that distorted information can occur either explicitly or implicitly. For example, stating a falsehood may be different in the eyes of some legal standards than applying a biased filter to express only a subset of the relevant facts. Expression that hurts another person may do so directly (i.e. a sin of commission) or indirectly (i.e. a sin of omission). A sin of omission occurs when an agent strategically or intentionally hides a fact that would prevent the other person from being hurt while only transmitting facts that cause harm—although the damaging facts are true. A natural question is whether expression that is an incomplete statement of truths (which includes selective expression resulting from a biased filter on the private information transmitted by the expresser) should be included in the definition of defamation. In other words, could there be a rationale for including expression that is not false in the definition of defamation? Our model demonstrates that restricting biased expression, where some information is deliberately suppressed, can be rationalized by our criterion of social welfare based solely on accuracy of subjective beliefs.

This equivalence of social harms caused by false statements and omitted truths has an application to the ‘actual malice’ rule established in *New York Times Co.* v. *Sullivan* (1964). According to this rule, ‘actual malice’ in US law is defined as knowledge of falsity or reckless disregard for the truth. In *New*

principle,” now referred to as the Harm Principle, which states that: “the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others’ (Mill, 1978). If we accept the harm principle, then the question arises: ‘What types of speech, if any, cause harm?’ The example Mill refers to involves dealers of corn. Mill suggests that it is acceptable to claim that corn dealers starve the poor, if such a view is expressed through the medium of the printed page. It is not acceptable, however, if those same words are delivered orally to an angry mob gathered outside the house of the corn dealer. The difference between these two cases is context. The context of speech when situated in front of the angry mob causes expression to, in Mill’s words, ‘constitute… a positive instigation to some mischievous act’, namely, to place the property rights, and possibly the life, of the corn dealer in danger. The high risk of inciting a riot justifies intervention. Of course, one could argue that the former constitutes ‘indirect’ harm, whereas the latter constitutes direct harm.

[6] Arrow (1997) criticized Radin by arguing that ‘the analogy of the democratic process with scientific progress certainly calls for freedom of inquiry and dissemination of its results’. Arrow perceived a contradiction in the chain of logic used to arrive at Radin’s conclusion given that, among the evidence Radin presented, were quotes from Dewey (1925) stating that democracy is the scientific method applied to social problems.

[7] According to Japanese civil law and criminal law, an expressed opinion does not constitute defamation if: (i) that expressed opinion concerns issues relevant to the ‘public interest;’ (ii) the purpose of expressing it corresponds with the public interest; and (iii) the expression is based on truth. For more details of Japanese defamation law, see Beer (1984). Article 310 of Korean criminal law specifies a similar standard, while Korean civil law provides that an expression may constitute defamation even if it is based on truth. US tort law is widely interpreted as allowing that an expression does not result in legal liability due to defamation if it is based on truth, which is different from Korean tort law.

[8] In USA, a statement does not need to be literally true in order for the defence against libel to be effective. US courts require that the statement is substantially true in order for the defence to apply it. It follows that even if the defendant states some facts that are objectively false, the defendant can nevertheless rely on that defence so long as the gist of the communication is true.

York Times v. *Sullivan*, the US Supreme Court ruled that the existing common-law definition of defamation violated the guarantee of free speech under the First Amendment of the Constitution. The solution adopted was to do away with the common-law presumptions of falsity and malice and place the burden on the plaintiff to prove that, at the time the defamatory statements were made, the defendant either knew them to be false or was reckless as to whether they were or not. Aside from many criticisms against the 'actual malice' rule,[9] our analysis challenges this rule from the viewpoint of Benthamite social efficiency.

In this article, we apply rational choice theory to analyse restrictions on free expression under the assumption that all individuals are rational in the sense that they: (i) make decisions by maximizing private benefits net of private costs, and (ii) update their beliefs about the state of the world using all available information according to Bayes' Rule. While writing this article, we found an alternative non-Bayesian approach addressing a similar issue. In particular, Glaeser and Sunstein (2014) use the term 'asymmetric Bayesianism' to explain the empirical finding that expressions or presentations of balanced information result in polarization, with the implication that 'more speech' does not help correct falsehoods. They argue that asymmetric Bayesianism leads those receiving a message whose beliefs are supported by the message to rationally believe it is true, while those whose beliefs conflict with the message discard it by believing that it is false. In their model of asymmetric Bayesianism, the same message can lead individuals to update their beliefs in different directions rather than converging to objective truth. They argue that this polarization problem can be resolved by introducing 'surprising validators' that can be considered credible. In this article, we argue that surprising validators can also backfire if we consider potential risks of strategic messaging.

The article is organized as follows. Section 2 sets up the model. Section 3 presents analysis using the model to draw policy implications regarding laws restricting free expression. Section 4 contains some discussions of modifications of the basic model provided in Section 2. In particular, we consider various liability rules and show that the anti-defamation law based on the negligence rule can lead to the socially optimal level of expressions by adjusting the due care level appropriately. Concluding remarks follow in Section 5.

2. Model

A society is characterized by the underlying state of the world denoted θ, which is assumed to be a fixed parameter unknown to each individual. At most, an individual may possess noisy information about the true state of the world.

An individual (referred to as Player 1) who has access to noisy information about θ can express it to the public. The public (Player 2) receives information expressed by Player 1. If Player 1's expressed information involves information about some third player (referred to as Player 3), then Player 2 can be interpreted as taking on this role, too. We assume that it is in the public's interest for all members of society to have more precise information about θ rather than less (i.e. social welfare is a decreasing function of the distance between subjective beliefs about θ and its true but unknown value).

[9] This rule has been criticized not only by plaintiffs but also by defendants. Plaintiffs complain that it is extremely difficult to prove 'actual malice', making it is nearly impossible to win a defamation suit; while defendants in the mass communications industry argue that the present system encourages frivolous lawsuits that cost substantial amounts to defend. Cass Sunstein (2014) commented that 'while it has granted the indispensable breathing space for speakers, it has also created a continuing problem for public civility and for democratic self-government'.

A representative individual who has noisy information contemplates whether to express it to the public. The following notation will be used throughout the article:

$b =$ gross private benefit from expressing information ($b \geq 0$);

$G(b) =$ probability distribution of b ($0 \leq G \leq 1$ for all $b \geq 0$);

$c =$ private cost of expressing information ($c > 0$);

$B =$ gross social benefit from the expression ($B \geq 0$);[10]

$C =$ social cost from the expression ($C \geq 0$);

$x =$ noisy signals about the true state of the world, θ, which are private to the expresser but verifiable by a regulator at a later stage of the game;

$m =$ expressed message.

We assume that $x = \theta + \epsilon_x$, where ϵ_x is symmetrically distributed around 0, with mean 0, $\text{var}(\epsilon_x) = \sigma_x^2 (> 0)$, which is re-parameterized as the precision parameter $h_x\left(= 1/\sigma_x^2\right) < \infty$.[11] Also, we assume that an individual's perceived benefit b is independent of m. That is, b does not change as a function of the agent's decision of which information to transmit. The perceived benefit of expression is a random draw from a set of possible valuations.

The Benthamite utilitarian social welfare function can be defined by:

$$W = (b + B) - (c + C).$$

If the society values freedom of speech itself rather than indirectly through its provision of more precise information, we could use a social welfare function which is the weighted sum of W and an additional term (V) measuring the social value added from enjoying the liberty of free speech:

$$\tilde{W} = (1 - \mu)W + \mu V,$$

where $\mu \in (0, 1)$ is the weight on the liberty and V is the intrinsic value of the free speech itself (net of any informational gain from exercising free speech).

3. Analysis

We first analyse the case in which only one signal is available. The subsequent subsection addresses the case in which multiple signals are available. The utilitarian social welfare function is assumed unless stated otherwise (in an extension with non-Benthamite social welfare functions considered below).

Suppose the potential expresser receives only one noisy information endowment x. If she is to express this information by choosing to transmit a message m, then all other agents' utilities who receive the expression are affected by B. If m is distorted (exaggerated) from her private information and other agents know the possibility of distortion (exaggeration) with certainty, B can be interpreted as the value accruing purely from the information inferred by discounting the face value m regardless

[10] By 'social benefit', we refer exclusively to the public's change in social welfare excluding any change in payoff by the individual who expresses information. Thus, b is not included in B.

[11] All mathematical expectations in this article take θ to be an unknown fixed value and not a random variable.

of m. However, if truth or falsity of a message is uncertain, it incurs social losses. The more uncertain the reliability of the message is, the higher social costs it incurs. Thus, the social cost associated with receiving message m, denoted by C, can be specified as a function that is proportional to its mean squared error (i.e. the expected squared distance from the objective truth): $C = F(MSE(m))$ where $F' > 0$ and $MSE(m) = E(m - \theta)^2$. This functional form captures the idea that people care about knowing the true value of θ. For simplicity, we assume that $F(MSE(m)) = \gamma MSE(m)$ where $\gamma > 0$.

Note that C depends on m while B does not. This simplifying assumption can be justified by the interpretation that the social benefit from expression—in the special case that it conveys perfectly accurate information—is represented by the parameter B and that social welfare decreases from this gross potential social benefit (B) as the result of any uncertainty of possible misinformation (measured by the mean squared error of the expressed information).[12] That is, all the decreases in the social welfare due to information distortion are assumed to be captured in C.

If $m = x$ (i.e. the expressed message is based on truth in the sense that it is a face-value transmission of received noisy information with no distortion applied), then $E(m - \theta)^2 = E(x - \theta)^2 = E(\epsilon^2) = \sigma_x^2 = 1/h_x$. This implies that disutility of receiving expressed messages is decreasing in the precision of the expresser's information.

Suppose that the expresser, for some reason,[13] distorts x in a systematic way, for example, $m' = x + \alpha$, where $\alpha \neq 0$. Here, α could be interpreted as a false distortion. The associated disutility is then $E(m' - \theta)^2 = E(x + \alpha - \theta)^2 = E(\alpha + \epsilon)^2 = \alpha^2 + 1/h_x$. It is easy to see that $MSE(m) = E(m - \theta)^2 < E(m' - \theta)^2 = MSE(m')$, because $\alpha^2 > 0$. Even if people recognize the possibility of distortion, they cannot correct their views by discounting m' unless they know the exact value of α. The baseline motivation for regulating free expression comes from this observation.

As mentioned earlier, θ may contain the information about someone else (Player 3) other than the receiver to whom the information was directly expressed (Player 2). In this case, distorted expression may ruin Player 3's reputation in the eyes of Player 2 by making $MSE(m')$ large. Such false information about someone and thereby damaging his reputation incurs a large social cost as well as the private cost to Player 3.[14] Thus, whether the cost is considered as private or social to the public, such defamation could be discouraged by liability or fines or both. Note that both liability and fines are effective means to correct the negative externalities an expression can generate to others.

In a deregulated environment without laws regulating free expression, a potential expresser expresses her information whenever benefits exceed costs, $b \geq c$. The measure of the mass of information that is expressed rationally is $\int_c^\infty dG(b) = 1 - G(c)$. The decision to express messages generates externalities, however, which affect social welfare. By the utilitarian criterion of social welfare, it is therefore socially optimal for an individual to express her information if and only if $b + B \geq c + \gamma MSE(m)$, or equivalently, $\tilde{b} \equiv b - c \geq \gamma MSE(m) - B$, i.e. her private net benefit from expression exceeds the social net cost. If the social benefit and the social cost coincide (i.e. if $B = \gamma MSE(m)$), then it is socially optimal to give each individual unrestricted freedom of expression. If $B \neq \gamma MSE(m)$, however, social optimum is not guaranteed and there may be either over-expression

[12] For example, in the 'Mad Cow' case which is elaborated in subsection 4.2, the report of MBC itself raised the issue of safety of imported beef, which has a social value (B) regardless of its truth. If the report were not true, then the social value would be discounted proportionally, depending on the magnitude of distortion (i.e. the extent to which the report was false).

[13] In our model, the expresser has no clear motive to distort information because her preference is independent from m. Since the expresser is by assumption indifferent over messages, there is also no compelling reason to exclude the possibility of transmitting distorted information. We will discuss the case in which the expresser prefers a particular value of m below.

[14] In our model, both costs are captured in $\gamma MSE(m)$.

or under-expression. In either case, there is a potential social-welfare motive for regulation: if $B < \gamma MSE(m)$, then too much expression may occur if the social net cost exceeds her net benefit from expression, and the social welfare criterion can rationalize restricting the expression of information.[15] The social-welfare motive for regulating speech is an increasing function of $MSE(m)$, or, equivalently, a decreasing function of precision h_x.

Under the no-liability rule whereby the expresser is never liable for any harm she causes, it is clear that there will be too much distorted expression. However, the social optimum can be obtained if the expresser is made liable for $\gamma MSE(m) - B$ whenever $B < \gamma MSE(m)$. Alternatively, $\gamma MSE(m) - B$ could be interpreted as a fine. As usual, social welfare maximization requires that the penalty for transmitting messages that hurt others be set equal to the negative externalities it generates, perfectly internalizing the negative externalities from free expression, insofar as the law enforcement is perfect.

This perfect internalization of negative externalities by setting liability equal to the marginal social cost of expression can be interpreted in terms of various liability rules. Firstly, it has the interpretation as the negligence rule. We can say that the expresser is negligent if and only if $\tilde{b} > 0$ (so she expressed something) but $\tilde{b} + B - C < 0$ i.e. $\gamma MSE(m) > B + \tilde{b}$. If an expresser is negligent, he should be liable for the damages $\gamma MSE(m) - B$ under the negligence rule.[16] It is well known that the negligence rule can achieve the social optimum. On the other hand, if $\tilde{b} < 0$ but $\tilde{b} + B - \gamma MSE(m) > 0$, then too little expression occurs and a social-welfare-maximizing government might seek to encourage more expression.[17]

Some might suggest a strict liability law in the defamation rule which would require that the expresser is liable whenever $B < C$. If an expresser were liable for any C such that $C > B$, then the rule would wind up restricting socially beneficial speech because under such a rule, expression is regulated even when $\tilde{b} > 0$ and $\tilde{b} + B - C > 0$.[18] For implications of the model based solely on the normative criterion of social welfare, we consider the following cases.

Distorted Expression: If $\alpha \neq 0$, then $m = x \quad \alpha$ could be interpreted as false information. This formalization shows one condition under which false expression is likely to cause social harms that exceed social benefits, thereby providing a rationalization of anti-defamation laws based on the falsity of expressed messages. In the context of the model and its specification of social welfare, a restriction on free expression can be rationalized as follows.

Claim 1: *It is socially rational to restrict expression m that defames or otherwise distorts information if and only if $\gamma MSE(x) < \tilde{b} + B < \gamma MSE(m)$, where $m = x \quad \alpha$ for some $\alpha \neq 0$.*

[15] We fully acknowledge that a more general, non-Benthamite social welfare function could easily lead to different conclusions about regulating or restricting speech. For example, if we assumed a fixed social-welfare cost anytime the principle of free expression is restricted, then of course this could be included in our Benthamite social welfare function and, consequently, negative externalities from free expression would have to be of a sufficiently large magnitude) to rationalize any restrictions on speech. Another possibility to consider is the risk that a current or future government might abuse restrictions on speech or perhaps disseminate biased information (e.g. propaganda or outright lies). The social welfare function could, once again, be modified to reflect such risks, which would imply, once again a more demanding threshold condition to guarantee that negative externalities avoided thanks to restricting speech would fully offset the expected costs (e.g. from abuses of such policies).

[16] See subsection 4.4 for an alternative interpretation of due care and negligence.

[17] For more elaborations, see Footnote 18.

[18] It is also well known that, similarly to the negligence rule, the strict liability rule achieves the social optimum. This is only true, however, when the potential injurer's action (e.g. expression) does not generate any positive externalities. If expression generates positive externalities such as B in our model, then the first-best result is not guaranteed by the strict liability rule (Kim, 2006). This is also the case for the negligence rule when $B > \gamma MSE(m)$.

Ungrounded Expression: Is it reasonable to restrict expression, even if the expressed message is true (i.e. $\alpha = 0$, interpreted here as face-value transmission of private information, which includes its noise term, rather than the objective truth about the state of the world, which is unknown to the expresser)? If negative externalities are only the result of imprecise information (i.e. the fact that precision h_x is small), then can there be a rationale for restricting speech?

To illustrate, consider the following defamation case. In the Korean presidential race in 2007, Bong-Joo Chung (a former Congressman not running in the presidential race) criticized opposition-party candidate Myung-Bak Lee. Chung claimed that Lee was a hidden owner of an investment consulting company previously found to have forged stock prices. Although Chung's allegation was based on a number of verifiable facts at the time, the court found Chung guilty of defamation because his sources were of uncertain veracity.

Our model translates that question into an inequality condition on social costs and private benefits: if an expressed message is based on verifiable information but the source is unreliable and the expresser knows it, i.e. knows that his or her information might be (harmfully) wrong, it should be subject to laws restricting defamation. In the preceding case, although Chung's allegation was based on some facts, $m = x$, (although not all facts were known: $m \neq \theta$), the court found Chung guilty of defamation on the grounds that h_x was small.[19]

Claim 2: *It is socially rational to restrict expression* $m = x$ *if and only if* $\tilde{b} + B < \gamma MSE(m)$, *where* $m = x$.

This claim is consistent with our argument that we used to show that the negligence rule can be socially optimal. More rationale for our claim that expression based on true but ungrounded information can be subject to defamation is provided in footnote 30 in subsection 4.4.

We observe that the possibility of rationalizing regulation of free speech (especially when said speech consists of messages that are not deliberate falsifications of private information) points to uncomfortable tension between the extent to which classical principles of liberty collide with the criterion of social welfare. The authors of this article do not agree on which of these normative principles should guide policies that regulate speech. The value of the model is that it clearly organizes these normative claims and makes clear that views about proper regulation of speech correspond to different relative weights or aggregations of value inherent in arguments based on political liberty versus social welfare maximization (i.e. minimizing harm from negative externalities generated by free expression). Those arguing that principles of individual liberty should guide policy are essentially applying a lexicographic rule or adding a negative term to the social welfare function that deducts social value whenever the principle is violated, or simply assuming that the social benefit from free expression and resulting diffusion of information B is so large that it exceeds $\gamma MSE(m)$ for any m. For observers that place weight on both normative principles (individual liberty and harm minimization in the social welfare framework), our analysis would imply that an appropriate large threshold is needed to quantify the reduction in social harm which offsets the cost of sacrificed individual freedom (if such a finite number exists). This point is revisited in subsection 4.5.

[19] In the US, some states stipulate that falsity is an element of defamation and that any plaintiff must prove falsity in order to recover damages. And even where falsity is not formally a requirement by defamation law, truth generally serves as an effective defence against accusations of libel or slander.

Before we close this section, we will briefly discuss the case that the private benefit of an individual is directly affected by the expression m. Suppose each individual has her own position τ and that her benefit from expressing m becomes smaller, the farther away it is from her position (favourite expression), where τ is uniformly distributed over [0, 1] among all possible individuals *à la* Hotelling (1929). Thus, if she transmits the expression m, then her private benefit is $b - t|m - \tau|$, where $t(> 0)$ is the unit traveling cost. (Imagine that an individual must travel the distance $|m - \tau|$ from her original position, τ, to express m.) If this individual is located at $\tau = 0$ (e.g. left-wing position), then her private benefit from expressing m is $b - tm$. Therefore, she would always have an incentive to choose m as low as possible in the absence of any penalty for distorting information. Likewise, if she is located at $\tau = 1$ (e.g. right-wing position), her private benefit is $b - t(1 - m)$, which is increasing in m, implying that she would have an incentive to exaggerate m as much as possible. In other words, each individual would choose to express his or her ideal location rather than a veridical or face-value transmission of the private information this individual observes (x). As a result, at least in the context of our model in which the social value of transmitting information is based on the objective accuracy of beliefs about the true state of the world, the expression of exaggerated messages would be of no use in the sense that no information or, at most, only misleading information is contained in them; therefore, the exaggerated messages as just described do not enable anyone to achieve more accurate beliefs about θ. Moreover, no one who receives such expressions would have any reason to believe them, because they only reveal the expresser's own position, τ, and do not contain the expresser's private information, x.[20] Because of this outcome in which all messages are uninformative and universally ignored, we stick to our assumption that b is independent of m.[21]

4. Discussions

In this section, we discuss possible extensions of our model to glean enriched implications from the model.

4.1 More than one expresser

One of the strongest rationales for free expression is that open dispute eventually leads to the truth or, at least, improves approximations to it, as Arrow (1997) said. Our intuition also seems to suggest that the aggregation of multiple independent messages transmitted by many expressers can reveal more precise information as the number of unrestricted expressions increases.

[20] If the private cost of distortion is proportional to the magnitude of distortion $m - x$ (e.g. due to an increase in the likelihood of penalty, or the psychic cost of a guilty conscience), then m may not be too far from x and would then take on a value somewhere between τ and x. In this case, the public could infer x from m if τ is known. So, an expresser's informational distortion accompanied by the pubic's discounting the message may give higher social welfare than regulating expression to induce truthful expression. At least, in this equilibrium, an expresser expresses what he wants and, as a result of discounting, there will be no information loss. However, if the inference is not perfect due to the lack of knowledge about the position τ or limited intellectual capability, an expresser will take advantage of imperfect rationality to decide what to express and thus an expression will reveal partial information about x and τ. Modelling the process underlying this stochastic inference is beyond the scope of our article and no further analysis in this direction is presented here.

[21] As far as b is independent of m, an expresser does not strictly prefer one message to another, that is, she is indifferent between distorting information and not distorting it. Since distortion can also occur in equilibrium, restriction of free expression can be socially productive.

To confirm our intuition, suppose that there are n expressers who have noisy signals about θ, $x_i = \theta + \epsilon_{x_i}$ where ϵ_{x_i}'s are independent and symmetrically distributed around 0, with mean 0, var $\left(\epsilon_{x_i}\right) = \sigma^2_{x_i}\left(= 1/h_{x_i}\right) > 0$ for $i = 1,\ldots, n$. For simplicity, we assume that $\sigma^2_{x_i} = \sigma^2_x$, i.e. $h_{x_i} = h_x$ for all i.

If all the expressers decide to express their own information without distortion (i.e. $m_i = x_i$), then the aggregated message that the receiver gets, $m = \sum_{i=1}^n m_i/n = \sum_{i=1}^n x_i/n$, gives the following mean squared errors, which clearly converges to zero as n becomes large:

$$\begin{aligned} MSE(m) &= E(m-\theta)^2 \\ &= E\left(\frac{\sum x_i}{n} - \theta\right)^2 \\ &= \frac{\sigma_x^2}{n}. \end{aligned}$$

Convergence to zero would seem to support the efficiency of free expression.

However, if expressers express information in a distorted way (i.e. choosing the expression $m_i = x_i + \alpha_i$), then the resulting mean square error is computed as follows:

$$\begin{aligned} MSE(m) &= E(m-\theta)^2 \\ &= E\left[\frac{\sum(x_i+\alpha_i)}{n} - \theta\right]^2 \\ &= \frac{1}{n^2}E\left[\sum_1^n (x_i + \alpha_i - \theta)\right]^2 \\ &= \frac{1}{n^2}E\left[\sum_1^n (\alpha_i + \epsilon_i)\right]^2 \\ &= \frac{1}{n^2}E\left[\sum_1^n \epsilon_i + A\right]^2 \\ &= \frac{\sigma_x^2}{n} + \frac{A^2}{n^2}, \end{aligned}$$

where $A = \sum_{i=1}^n \alpha_i$. If $\alpha_i = \alpha$ for all i, we have $MSE(m) = \frac{\sigma_x^2}{n} + \alpha^2$ which clearly does not converge to zero as $n \to \infty$. This result implies that free dispute, left unrestricted, does not necessarily guarantee that it leads to the truth, insofar as a certain proportion of expressers distort information. What if we can expect a self-correction mechanism of public discourse, i.e. the possibility that distorted messages might be possibly corrected by some others? We could model this possibility by interpreting α_i as a random variable which is independent of ϵ_i and has identical and independent distributions with mean 0 and var$(\alpha_i) = \sigma^2_\alpha(> 0)$. That is, there might be a message $-\alpha_i(< 0)$ that corrects $\alpha_i(> 0)$. Then, the mean square error is

$$\begin{aligned} MSE(m) &= E(m-\theta)^2 \\ &= \frac{1}{n^2}E\left[\sum_{1}^{n}(\epsilon_i+\alpha_i)\right]^2 \\ &= \frac{\sigma_x^2}{n}+\frac{\sigma_\alpha^2}{n}, \end{aligned}$$

since ϵ_i and α_i are independent. This converges to zero as $n\rightarrow\infty$. To summarize, we have:

Claim 3: *As the number of expressers grows to infinity, free expression leads to the truth if all expressers express their own information. If expressers distort their information, however, it does not necessarily reveal the truth, but it leads to the truth if the error-correction mechanism of pubic discourse is strong enough.*

This claim suggests that whether public discourse leads to the truth or not depends on how well the error-correction mechanism functions in the society. We assumed above that the mechanism works perfectly by assuming that the distribution of α_i is not skewed, i.e. symmetric around 0. However, is public discourse's power of error correction strong enough in reality? We do not want to give a definite answer for the question here, but at least we can assert that public discourse and the court are two complementary error-correcting mechanisms of public expressions and that the court is the last recourse for error correction after a correction process of public discourse may fail.[22] In the remaining analysis, we consider only models with one potential expresser.

4.2 Multiple signals

In 2008, the Korean television network MBC aired an episode of its show *PD Notebook* titled, 'Is U.S. Beef Really Safe from Mad Cow Disease?' MBC was then charged with slandering public officials in the Korean government by exaggerating the risk of Mad Cow Disease. The programme stated that an American woman's cause of death was a variant of Mad Cow disease. Whereas US media presented various possibilities for her cause of death, PD Notebook was alleged to have deliberately omitted those other possibilities, selectively focusing only on the brain-wasting disease. A Korean court ordered MBC to air a correction affirming the government's claim of slander.

As MBC obtained many signals about the safety of US beef in this case, an expresser may acquire not simply a single signal but multiple signals. Let x and y represent two noisy signals available to a potential expresser. We assume that $x=\theta+\epsilon_x$ and $y=\theta+\epsilon_y$, where ϵ_x and ϵ_y are independent and symmetrically distributed around 0, with mean 0, $\text{var}(\epsilon_x)=\sigma_x^2(=1/h_x)>0$, $\text{var}(\epsilon_y)=\sigma_y^2(=1/h_y)>0$, respectively. Again, h_y is the precision parameter of signal y. Let λ_1 denote the

[22] In 2011, Tablo, a famous Korean singer, filed the libel suit against 22 bloggers for disseminating libelous disinformation questioning his academic achievements. The rumour about Tablo's education career started in 2010 from several bloggers who claimed that the singer falsely promoted himself as a Stanford graduate, and the false information was spread in cyberspace. At that time, an Internet Cafe 'Let's Ask Tablo the Truth' with more than 200,000 members was opened and they kept producing false claims and suspicions. Whenever evidence supporting Tablo was presented, more and more allegations against him followed. For example, even if Tablo posted his diploma on the Internet, they claimed that it was fake for several specious reasons. Even after Thomas Black, an associate vice provost of student affairs at Stanford verified his academic record, it was overshadowed by their claim that he was not a reliable professor. The scandal was finally terminated in 2011 by Tablo's indictment of the bloggers on charges of defamation, but he had to suffer a lot in the meantime.

weight that Player 1 places on x when choosing messages that are chosen from the set of convex combinations of the two available signals, as follows. Player 1's expressed message is $m = \lambda_1 x + (1 - \lambda_1)y$, where $\lambda_1 \in [0, 1]$. To compute its mean squared error (MSE), we have:

$$\begin{aligned} MSE(m) &= E(m - \theta)^2 \\ &= E(\lambda_1 x + (1 - \lambda_1)y - \theta)^2 \\ &= E[\lambda_1(x - \theta) + (1 - \lambda_1)(y - \theta)]^2 \\ &= \lambda_1^2 E(x - \theta)^2 + \lambda_1(1 - \lambda_1)E(x - \theta)(y - \theta) + (1 - \lambda_1)^2 E(y - \theta)^2 \\ &= \lambda_1^2 E(\epsilon_x^2) + (1 - \lambda_1)^2 E\left(\epsilon_y^2\right) \\ &= \lambda_1^2 \sigma_x^2 + (1 - \lambda_1)^2 \sigma_y^2 \\ &= \left(\sigma_x^2 + \sigma_y^2\right)\lambda_1^2 - 2\sigma_y^2 \lambda_1 + \sigma_y^2. \end{aligned}$$

The $MSE(m)$ associated with m achieves its minimum at $\lambda_1 = \frac{\sigma_y^2}{\sigma_x^2 + \sigma_y^2} = \frac{h_x}{h_x + h_y}$. The minimum value of $MSE(m)$ is $\frac{1}{h_x + h_y}$, which is smaller than $\frac{1}{h_x}$ and $\frac{1}{h_y}$. If $h_x = h_y = h$, then the minimum MSE is achieved by the unbiased expression $m^* = \frac{x+y}{2}$ which is the optimal expression. Because these two signals have equal precision, the MSE of m^* is reduced by half. Stated equivalently, by basing a message on two signals with independent error terms ($m^* = \frac{x+y}{2}$) instead of one signal ($m = x$ or $m = y$, as in the previous subsection), the precision of the expressed information increases by a factor of two.

We can interpret $\lambda_1 = 1$ as the decision to selectively express only one (signal x) of the two available signals. The decision $\lambda_1 = 0$ corresponds to selective expression of y only. If $\lambda_1 > \frac{h_x}{h_x + h_y}$, then we can further interpret the expression as exaggerating the importance of x relative to y. Similarly, $\lambda_1 < \frac{h_x}{h_x + h_y}$ corresponds to exaggerating the importance of y. The above calculation of MSE associated with the message $m = \lambda_1 x + (1 - \lambda_1)y$ shows how exaggeration or linear distortions of any message (within the family of convex combinations of x and y) generates social harm in our social welfare analysis based on the principle of valuing dissemination of true information. In particular, when $\lambda_1 = 0$ or 1 (i.e. the expresser intentionally suppresses one signal), then after the values of x and y are realized *ex post*, the resulting expression will be maximally inefficient as $MSE(m)$ achieves its maximum, assuming that $h_x = h_y$. Therefore, we have:

Claim 4: *It is socially optimal to regulate the expressed message m as defamation whenever* $\gamma MSE(m^*) < \tilde{b} + B < \gamma MSE(m)$ *for* $m \neq m^*$.

This result has a further implication for the 'actual malice' rule established in *New York Times Co.* v. *Sullivan* (1964). As mentioned earlier, 'actual malice' in US law is defined as knowledge of falsity or reckless disregard for the truth. In *New York Times* v. *Sullivan*, the US Supreme Court overruled a State court in Alabama that had found New York Times Co. guilty of libel. Even though some of what New York Times Co. printed was false, the Court ruled in its favour, saying that libel of a public official requires proof of 'actual malice', which was defined as a 'knowing or reckless disregard for the truth'. Again, the social welfare analysis seems to be consistent with both 'actual malice' and 'reckless disregard for the truth' (even if the expresser's statements do not contain false information) as grounds for libel.

4.3 When the receiver also has some information

If noisy information about the true state of the world is also available to the receiver (Player 2), then the receiver (of the message expressed by Player 1) will update beliefs by computing a weighted average of the receiver's private information and the expressed message received from Player 1. Denote the receiver's private information as z, while x continues to denote the expresser's (i.e. Player 1's) information. We assume that $z = \theta + \epsilon_z$, where ϵ_x and ϵ_z are independent, and ϵ_z is symmetrically distributed around 0 with mean 0 and precision $h_z (< h_x)$.

Consider the following weighted average $w = \lambda_2 m + (1 - \lambda_2)z$, where λ_2 is the weight that Player 2 places on Player 1's expression. (The receiver's choice of λ_2 represents the weight placed on the message received from an expresser). If, after observing m, the receiver believes that the expresser's message is an unbiased transmission based on a signal possessed by the expresser (i.e. $m = x$), then we have the following MSE calculation:

$$\begin{aligned} MSE(w) &= E[\lambda_2 m + (1 - \lambda_2)z - \theta]^2 \\ &= E[\lambda_2(x - \theta) + (1 - \lambda_2)(z - \theta)]^2 \\ &= E[\lambda_2 \epsilon_x + (1 - \lambda_2)\epsilon_z]^2 \\ &= \frac{\lambda_2^2}{h_x} + \frac{(1 - \lambda_2)^2}{h_z}. \end{aligned}$$

To minimize MSE, the receiver will choose the weight λ_2 to satisfy the following first-order condition:

$$\frac{\partial MSE(w)}{\partial \lambda_2} = 2\left[\lambda_2^*\left(\frac{1}{h_x} + \frac{1}{h_z}\right) - \frac{1}{h_z}\right] = 0. \tag{1}$$

The solution to the first-order condition given by equation (1) provides the MSE-minimizing weight that receivers apply to an expresser's message: $\lambda_2^* = \frac{h_x}{h_x + h_z}$. As h_x increases, $1 - \lambda_2^* = \frac{h_z}{h_x + h_z}$ decreases and approaches zero. In other words, the receiver's optimal weighting function can be interpreted as the receiver placing less weight or confidence on her own private information because she believes that the expresser is more reliable (i.e. the receiver believes that the expresser's information is more precise).

Given the receiver's optimal weight λ_2^* for integrating received messages together with the receiver's private information (under the assumption that the expresser is transmitting an unbiased copy of the expresser's private information), we next consider what happens if the expresser chooses to send a distorted message $m' = x + \alpha$, where $\alpha \neq 0$. The receiver's belief w' when receiving the biased message (while mistakenly applying optimal confidence weightings based on the assumption that the received message is unbiased) is $w' = \lambda_2^*(x + \alpha) + \left(1 - \lambda_2^*\right)z$, and its MSE is:

$$\begin{aligned} MSE(w') &= E[\lambda_2^*(x + \alpha) + \left(1 - \lambda_2^*\right)z - \theta\,]^2 \\ &= H(h_x)^2\left(\alpha^2 + \frac{1}{h_x}\right) + (1 - H(h_x))^2\left(\frac{1}{h_z}\right), \end{aligned}$$

where $H(h_x) = \frac{h_x}{h_x + h_z}$.

As intuition would suggest, $MSE(w')$ is unambiguously increasing in the expresser's bias α. Intuition also suggests that the negative externality (as measured by $MSE(w')$) caused by biased messages could become more severe whenever the objective and commonly known precision of the expresser's information h_x increases. If an expresser's private information is very precise, then receivers would be overconfident in the credibility of the message, placing too much weight on it, thus exposing themselves to greater damage from the biased message expressed by a highly credible expresser (i.e. someone known to possess very precise private information).

The effect of h_x on $MSE(w')$ is, in general, indeterminate. We will show, however, that the intuition just described—in which greater precision leads to greater social costs—is captured within a large and dense subset of the parameter space in our model. An increase in h_x has two effects: on the one hand, h_x reduces $MSE(w')$ because x becomes more informative; on the other hand, h_x has a positive (i.e. socially damaging) effect on $MSE(w')$ through the bias term because more weight is placed on distorted information. The net effect therefore depends on both the precision of the expresser's private information h_x and the magnitude of distortion α. Formally, we have:

$$\frac{\partial MSE(w')}{\partial h_x} = -\frac{H^2}{h_x^2} + 2H'\left[H\alpha^2 - (1-H)\frac{1}{h_z}\right], \tag{2}$$

where $H' = \partial H/\partial h_x > 0$. The first term corresponds to the socially beneficial effect (reducing $MSE(w')$), while the second bracketed term corresponds to the indirect and socially costly effect (possibly increasing $MSE(w')$) through the weight the receiver places on the message of the expresser.[23] Thus, the net effect depends on the precision level h_x and the magnitude of the distortion α.

We can see from the formula for $\frac{\partial MSE(w')}{\partial h_x}$ given BY equation (2) that if the distortion (α) is sufficiently large, then the socially damaging effect that increases $MSE(w')$ will dominate. Similarly, if h_x is very large (i.e. the expresser has much more precise information than receivers do), then it is easy to see that increased expertise or credibility of the expresser (i.e. an increase in h_x) is socially costly ($\frac{\partial MSE(w')}{\partial h_x} > 0$), because receivers mistakenly place too much weight on the expresser's message.[24] This result could be used to rationalize applying a stricter standard to messages expressed by socially influential public figures than to messages by anonymous individuals.[25] Information distortion by a more reliable expresser is riskier from a social welfare point of view. The upshot is that the standard to regulate the freedom of expression should not be determined solely on the basis of either α or h_x but from joint consideration of the relative magnitudes of α and h_x.

This argument also has an interesting implication regarding Glaeser and Sunstein's (2014) result that the polarization problem can be resolved by strategic messaging with surprising validators. They argue that even asymmetric Bayesian decision-makers cannot dismiss information that comes from

[23] Strictly speaking, the indirect effect itself consists of two effects of opposite sign. One (i.e. the first term inside the square bracket) is the socially costly effect of placing more weight on distorted information α; and the other (i.e. the second term inside the square bracket) is the socially beneficial effect of placing less weight on the receiver's less precise information z.

[24] At the opposite extreme when the expresser has very little information (i.e. h_x is close to zero), the sign of this effect once again depends on the magnitude of α relative to h_x, although in this case there is more scope for the sign to be negative, which would allow for socially beneficial increases in the precision of an expresser's information who begins from a very low base rate of precision. Across the entire range of precision parameters, the effect of more precision on social harm depends on the relative magnitudes of α and the exogenously given precision of private information as parameterized by h_z and h_x.

[25] If the regulation is made in such a way, the government may abuse its powers by regulating an influential figure's statements that it dislikes rather than risky statements. This may be a strong argument for why free speech should be protected constitutionally. The next footnote provides an example for this argument.

someone who is highly credible, no matter how unwelcome that information is. Our model suggests, however, that introducing surprising validators could make this strategy risky insofar as receivers trust the highly credible expresser too much.[26]

4.4 Efforts to ascertain the truth of expressed information

What if the expresser had the chance to invest some resources to check the truth of the signal she receives? Let e, $e \geq 0$, represent effort expended on information acquisition. Effort is assumed to be observable and verifiable. It is natural to assume that h_x is strictly increasing in e (i.e. $h_x{}'(e) > 0$). The more effort the potential expresser expends, the more accurate the signal she acquires is.

The social welfare function must then incorporate this additional term involving the cost of effort aimed at improving the precision of private information:

$$W(m, e) = b + B - c - \tilde{C}(m, e),$$

where $\tilde{C}(m, e) = \gamma MSE(m; e) + e$. We assume that the expresser makes her decision in two steps: she first chooses e, and then decides whether to express $m = x$ with precision $h_x(e)$. In this subsection, we assume, for simplicity, that there is no possibility for the expresser to distort the message, which implies that she makes a binary decision, either expressing $m = x$ or not expressing at all.

We analyse this two-step problem of the expresser by backward induction. Given e, it is socially optimal for an expresser to express $m = x$ if $h_x(e)$ is sufficiently large (or equivalently, $MSE(m)$ is sufficiently small) that:

$$\tilde{b} + B > \gamma MSE(m), \tag{3}$$

because e is already a sunk cost when the second-step binary decision (to express or not to express) is made. The social-welfare-maximizing decisions by the expresser depend on the value of e, denoted e^*, that minimizes (i.e. is the argmin) of the following cost term from the social welfare function:

$$\tilde{C}(m = x, e) = \gamma MSE(m) + e = \frac{\gamma}{h_x(e)} + e.$$

Social welfare is maximized when the expresser chooses $e = e^*$ and expresses $m = x$ if the following inequality holds:

$$\tilde{b} + B > \gamma MSE(m) + e^*; \tag{4}$$

and chooses $e = 0$ and no expression otherwise (See Fig. 1). Note that inequality (4) implies inequality (3). That is, an expresser expresses her information whenever she decides to ascertain its truth. If the social cost is lower than the sum of the private net benefit and the social benefit when the potential expresser does her best to reduce it, then it is socially optimal for her to express the information she obtains by taking the efficient level of effort e^*. On the other hand, if the social cost exceeds the sum of

[26] For example, it is widely believed that the global financial crisis was aggravated by Alan Greenspan, the former chairman of US Federal Reserve Board (FRB), because he underestimated the risk of mortgage-backed securities and interest rate derivatives. Another example is the case of a financial blogger posting under the username Minerva, who was arrested in 2009 for spreading false rumours that allegedly destabilized the South Korean economy. The arrest was controversial. Minerva posted articles that correctly predicted the collapse of Lehman Brothers and the sharp decline of the South Korean won. Minerva became one of the most influential critics of the government's policies and came to be known as the 'Internet Economic President'.

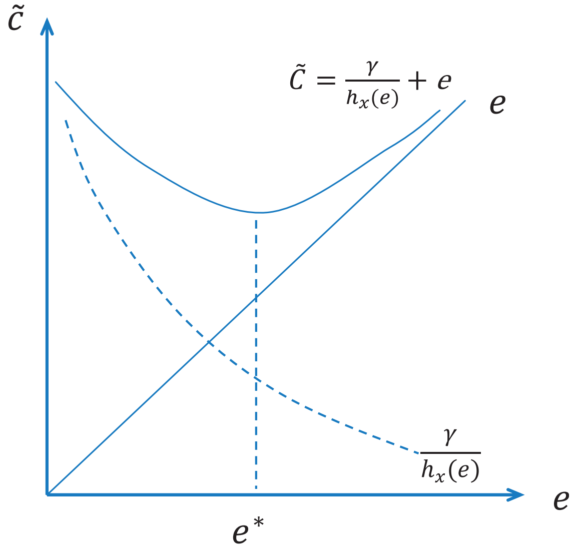

FIG. 1. Socially optimal level of effort.

the benefits, then it is socially efficient for her to choose zero effort acquiring better information and not to express anything at all (because the mean-squared error is large).

The incentive of the potential expresser to acquire information and to express the information she received depends on the liability rule. Under the no-liability rule, she expresses the information that she received with no effort as long as $b > c$, because she has no incentive to expend any effort to ascertain the truth of the information.

Under the strict liability rule whereby the expresser is liable for the social cost equal to $\gamma MSE(m)$, however, the expresser's objective given e is to maximize her utility:

$$W_S = \begin{cases} b - c - \gamma MSE(m; e) & \text{if she expresses } m \\ 0 & \text{if she expresses nothing.} \end{cases}$$

Therefore, given e, she expresses m if:

$$\tilde{b} > \gamma MSE(m; e), \tag{5}$$

and expresses nothing, otherwise. Thus, in the information acquisition stage, she will choose e to maximize $b - c - \tilde{C}$, or, equivalently, minimize $\tilde{C} = \gamma MSE(m; e) + e$. The minimum is obtained by the expresser choosing $e = e^*$ and $m = x$ whenever $W_S(e^*) > 0$, or, equivalently, if:

$$\tilde{b} > \gamma MSE(x) + e^*; \tag{6}$$

and the expresser chooses to not ascertain the truth and not express m, otherwise. Again, by the sunk cost argument, inequality (6) implies inequality (5).

Comparison of inequalities (4) and (6) implies that the expresser's private optimum under the strict liability rule is identical to the social optimum if $B = 0$. But more generally, the expresser's private optimum under strict liability need not be identical to the social optimum, because the expresser does not care about the social benefit B under the strict liability rule. This is due to the usual insight that the strict liability rule cannot lead to the first-best outcome when the activity of the agent (in this case, expression) generates both positive and negative externalities (Kim, 2006). Therefore, the strict liability rule deters too much expression.

Finally, consider the negligence rule.[27] If the due care threshold, $\overline{e}$, is set to the socially optimal level of care,[28] $e*$, then the expresser's utility under the negligence rule is:

$$W_N = \begin{cases} b - c - \gamma MSE(m, e) - e & \text{if } e < \overline{e} \text{ and she expresses } m \\ b - c - e & \text{if } e \geq \overline{e} \text{ and she expresses } m \\ -e & \text{if she expresses nothing.} \end{cases}$$

Note that $e \geq \overline{e}$ implies that $MSE(m; e) = 1/h_x(e) \leq 1/h_x(\overline{e}) = MSE(m; \overline{e}) \equiv MSE(m)$. Therefore, under this rule, the potential injurer (i.e. the expresser) is liable only if she fails to meet the due care standard (i.e. if $e < \overline{e}$, or, equivalently, if $MSE(m) = 1/h_x(e) > MSE(m)$).[29] Thus, she will choose $e = \overline{e} = e^*$ and $m = x$ if:

$$\tilde{b} > \overline{e}. \tag{7}$$

In other words, the expresser chooses the socially optimal level of due care and expresses the information she obtains[30] whenever her net private benefit from expression exceeds the cost of ascertaining the truth; but if this private cost exceeds the private benefit (i.e. she finds that taking effort $e = \overline{e}$ and expressing $m = x$ give lower utility than no expression at all, then she will choose $e = 0$.[31]

The government can use the due care threshold, $\overline{e}$, as a policy variable. To see this, consider two due care thresholds, $\overline{e}_1 = e^*$ and $\overline{e}_2 < e^*$, as shown in Fig. 2. If the due care threshold is $\overline{e}_1 \equiv e^*$, then expression m occurs whenever:

$$\tilde{b} > \gamma MSE(e^*) + e^*. \tag{8}$$

[27] Under US law, if someone makes a truthful statement based on false facts (i.e. $= x$, where h_x is very low), then a plaintiff can still win the case, especially if he or she can prove negligence in the collection of the false facts used to support the statement. This possibility would seem to support the negligence rule.

[28] The optimal standard is defined as the care level that maximizes social welfare W which is an addition of private and social net benefits in both models in Section 2 and this section. However, in the model of Section 2, the optimal care level was defined in terms of the mean square error of an expression, while, in this model, it can be defined in terms of the effort required to acquire more precise information e.

[29] We used a proxy $B + \tilde{b}$ as a possible interpretation for $MSE(m)$ in section 3 in which the truth-ascertaining activities were not available.

[30] This provides a rationale for regulating expression based on true but ungrounded information. Such a regulation would give a potential expresser the incentive to expend enough efforts to ascertain the truth of her information.

[31] It is the burden of the judge to prove whether the potential expresser took $e \geq \overline{e}$ or $e < \overline{e}$, but it is not clear who should determine whether a statement is true or not. In fact, the cases should be distinguished as to whether $e \geq \overline{e}$ and whether $m = \theta$. The former inequality is not about what the truth is, but about how much expertise is achieved. Therefore, the former inequality is determined independently of the latter inequality. Since the latter inequality requires a much higher degree of expertise than the former inequality, the judge would usually decide on the latter inequality with the help of experts. In the famous case of Woo Suk Hwang who was eventually found to have published faked research results on human stem cells, the allegations were first reported on the TV show *PD Notebook* by the broadcasting company MBC. When fans of Dr Hwang sued MBC for defamation, an investigation panel consisting of scientists in Seoul National University was assigned the task of determining whether claims made by MBC were true.

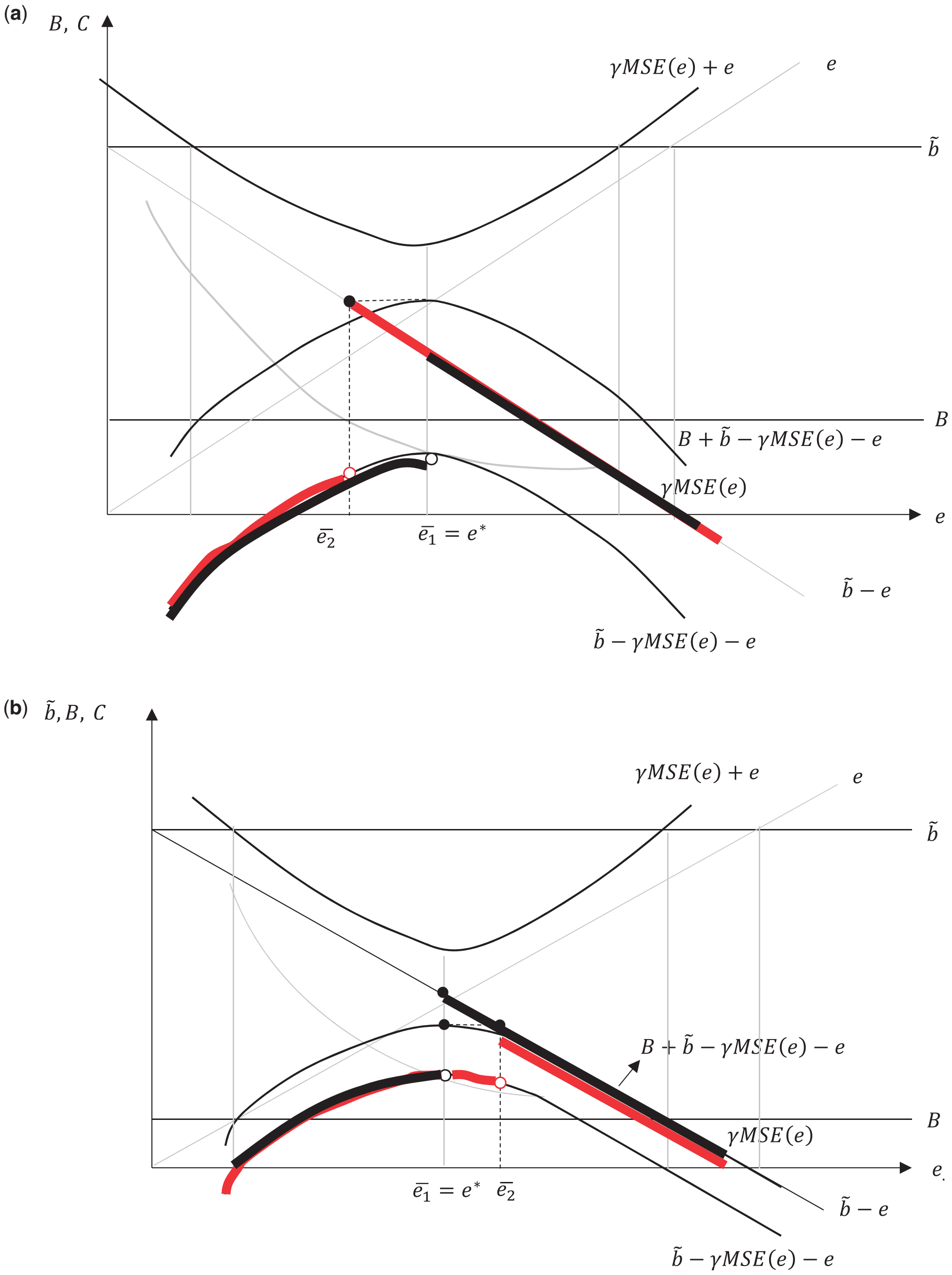

FIG. 2. The expresser's due cares under the negligence rule when (a) B > γMSE and (b) B < γMSE.

Because the social optimum requires that the expresser will actually express m if $\tilde{b} > \gamma MSE(m, e^*) - B + e^*$ based on inequality (4), one observes that too much expression (according to the model) occurs if $B < \gamma MSE(m, e^*)$, and too little expression occurs if $B > \gamma MSE(m, e^*)$.

The question naturally arises as to whether a more flexible threshold of due care, other than e^*, might eliminate these policy errors (i.e. too much or too little expression), thereby achieving the social optimum. First, consider the case of $B > \gamma MSE(m, e^*)$. If the due care threshold is $\bar{e}_2$, so that:

$$\tilde{b} - \bar{e}_2 = \tilde{b} + B - \gamma MSE(m) - e^*, \tag{9}$$

then $\bar{e}_2 = e^* + \gamma MSE(m) - B < e^*$. We can then show that the negligence rule achieves the social optimum, as shown in Fig. 2a. The left-hand side of equation (9) is the expresser's net benefit when she chooses the due care threshold $\bar{e}_2$, while the right-hand side is the social benefit when she chooses the socially optimal level of care. Under the negligence rule with the due care threshold e^*, the expresser does not have a strong incentive to express, because she does not care about the term, $B - \gamma MSE(> 0)$. By allowing a more lenient due care threshold, the negligence rule can lower the expresser's cost, thereby incentivizing her to express and expend effort ascertaining truth at the socially optimal level.

If $B < \gamma MSE(m, e^*)$, then we can find a due care threshold, $\bar{e}_2(> e^*)$, such that $\tilde{b} - \bar{e}_2 = \tilde{b} + B - \gamma MSE(m, e^*) - e^*$, which means that $\bar{e}_2 = e^* + \gamma MSE(m, e^*) - B(> e^*)$, because $\gamma MSE(m, e^*) - B > 0$. In this case ($B < \gamma MSE(m, e^*)$), the government requires a harsher threshold of due care to discourage excessive (i.e. socially damaging) expression and achieve the first-best outcome, as shown in Fig 2b. We will refer to the negligence rule with due care e^* as the 'specific negligence rule' and to the negligence rule with an arbitrary threshold of due care as a 'general negligence rule'.

Next, we compare outcomes under different liability rules. The conditions defining the first-best liability rule achieving the Social Optimum, the Strict Liability rule, and the Specific Negligence Rule, are summarized as follows:

$$\begin{aligned} &\text{(Social Optimum)} && \tilde{b} > \gamma MSE(m) + e^* - B; \\ &\text{(Strict Liability)} && \tilde{b} > \gamma MSE(m) + e^*; \\ &\text{(Specific Negligence Rule)} && \tilde{b} > e^*. \end{aligned}$$

Again, we will consider two situations, $B > \gamma MSE(m)$ and $B < \gamma MSE(m)$. If $B > \gamma MSE(m)$, we have the following order:

$$\gamma MSE(m) + e^* > e^* > \gamma MSE(m) + e^* - B. \tag{10}$$

The inequality above implies that expression is over-regulated under both the strict liability rule and under the negligence rule with due care e^*, because they lead to expression of m only when $\tilde{b}$ is much greater than the socially optimal level $\gamma MSE(m) + e^* - B$. On the other hand, if $\gamma MSE(m) > B$, we have:

$$\gamma MSE(m) + e^* > \gamma MSE(m) + e^* - B > e^*. \tag{11}$$

In both cases, the strict liability rule deters expression the most (i.e. is most chilling). But in the case that $\gamma MSE(m) > B$, then, interestingly, the policy maker (whose assessment of social damage is reflected in the inequality $\gamma MSE(m) > B$) views expression as under-regulated under the specific negligence rule but still over-regulated under the strict liability rule. This implies that the specific negligence rule does not have enough chilling effect from the viewpoint of the policy maker when the social benefit from the expression is regarded as small relative to its costs.[32] The intuitive reason for this is that it is socially optimal for an expresser not to express m if B is small (relative to $\gamma MSE(m)$); but under the specific negligence rule, the expresser does not take B into account and expresses m as long as private effort costs, e^*, are small. The specific negligence rule is basically a regulation requiring that expressers set their effort level, e, to e^*. If e is regulated, then $MSE(m, e)$ is also regulated, but B is not. Therefore, the specific negligence rule cannot regulate B, and so it may under-regulate expression depending on the size of B. Figure 3 summarizes this comparison.

Of course, if we consider a modified strict liability rule whereby the expresser is liable for $\gamma MSE(m) - B$ if $B < \gamma MSE(m)$ and receives a subsidy equal to $B - \gamma MSE(m)$ if $B > \gamma MSE(m)$, then the expresser's objective is to maximize her utility:

$$W_S = \begin{cases} b - c - (\gamma MSE(m, e) - B) - e & \text{if she expresses } m \\ -e & \text{if she expresses nothing.} \end{cases}$$

Maximizing the utility function above, the expresser chooses $e = e^*$ and $m = x$ only if $\tilde{b} + B - \gamma MSE(x) - e^* > 0$; and otherwise, she expresses nothing. This means that the expresser's private optimum under this modified strict liability is identical to the social optimum. The modified strict liability rule internalizes all the externalities that expression generates. Therefore, the usual insight applies to this first-best result except that the expresser is liable not for the social cost she incurs, but for the net cost, $\gamma MSE(m) - B$.[33]

4.5 Value of liberty

Multiple arguments have been put forward by social choice theorists (spanning a broad spectrum of points of view) in favour of valuing free expression intrinsically on non-consequentialist grounds (Sen, 1970, 1976; Pattanaik, 1994; Suzumura and Xu, 2001). Quite apart from any consequentialist payoffs derived from the exercise of the right to free speech, individuals may feel strongly that their wellbeing is improved by having available (e.g. strict meta-preferences in favour of choice sets that include) some elements that they in fact never intend to use (Sen, 1999; Suzumura and Xu, 2004; Berg and Gigerenzer, 2010; Dave and Dodds, 2012; Berg, 2014). Intrinsic individual and / or social valuation of the right to free speech would translate to our analysis as including in the social welfare function a strictly positive term measuring the intrinsic value of living in a society in which free expression is permitted (net of any informational effects from the exercise of free speech). If society values individual liberty itself, then the social welfare function that we should consider is $\tilde{W} = (1 - \mu)W + \mu V$ and the implication of the model is easily modified as predicted.

To illustrate, we consider only the case in which there is one signal available. Under the Benthamite social welfare function, we showed a social-welfare rationalization for regulation of distorted expression if and only if $\gamma MSE(m') > \tilde{b} + B > MSE(m)$. Now, under the modified social welfare function, the condition rationalizing free expression becomes:

$$\tilde{W} = (1 - \mu)\left[\tilde{b} + B - \gamma MSE(m')\right] + \mu V > 0,$$

[32] This corresponds with the argument of Miller and Perry (2013) that the fault requirement has the effect of mitigating the chilling effect.

[33] For a similar insight, see Kim (2006).

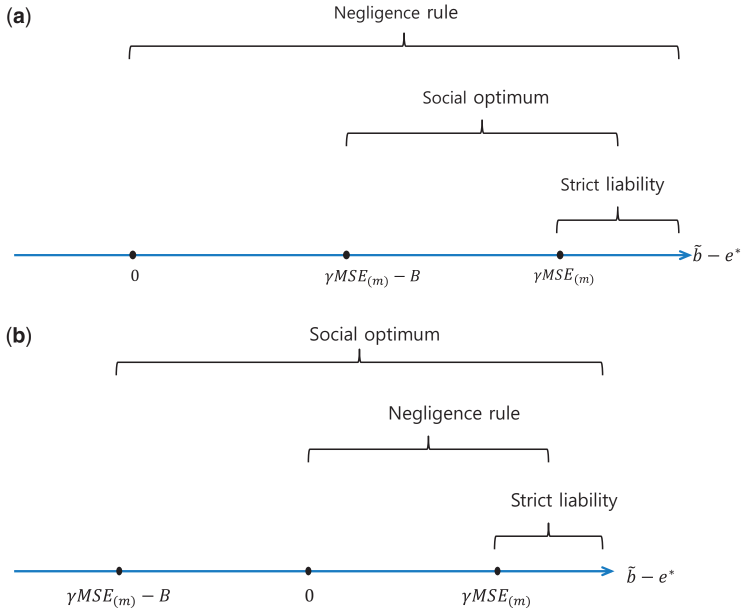

FIG. 3. Comparison of outcomes under various liability rules when (a) $\gamma\text{MSE(m)} > B$ and (b) $\gamma\text{MSE(m)} < B$.

or, equivalently, $\gamma MSE(m') - \left(\tilde{b} + B\right) < \frac{\mu}{1-\mu} V$. This condition rationalizing unrestricted speech may easily hold even if $\gamma MSE(m') > \tilde{b} + B$. Thus, if V or μ is very high, then expression should be left unregulated even though distortion $MSE(m')$ is large. In other words, as V or μ become greater, defamation laws should be applied more leniently. And as γ becomes larger (e.g. the information is considered to be more related to the public interest), then the model would suggest that these laws should be less leniently applied.

5. Conclusion

In this article, we examined under what circumstances expression subject to laws on libel and slander should be regulated based on weights allocated to informational efficiency and individual utilities of an expresser and those possibly harmed in a social welfare function framework. In this framework, regulation of expression can be justified when that expression is intentionally distorted or when the source itself is known by the expresser to be potentially fallible even if the expression itself is undistorted. We also showed that the negligence rule can lead to the socially optimal level of expression by adjusting the due care standard appropriately. Finally, we discussed how our results can be affected by alternative assumptions. For example, we showed that free competitive expression by many people can lead to the truth as the number of people gets larger if there is no intentional distortion in expression or if public discourse keeps correcting information distortion even if there is a possibility of intentional information distortion. Also, we stressed that our results depend on what is included in the social welfare function and its specification.

It is of course reasonable to ask what practical results follow from our theoretical exercise. Can it be applied to real cases? Richard Posner (1986) proposed the free speech formula that Hand used in the 1950 case, *United States* v. *Dennis*.[34] The formula, which is a counterpart of Hand's negligence formula, requires a court to determine the constitutionality of a regulation that limits freedom of speech by asking 'whether the gravity of an evil justifies such invasion of free speech'. Translated to an inequality condition, Posner's formula requires the court to regulate if and only if $V < PL$, where V is the lost benefit caused by the regulation (i.e. any loss from suppressing otherwise valuable information), P is the probability that the speech will do harm, and L is the social cost of the harm. There is an obvious analogy between the Hand formula and our social welfare analysis. In our formula given in equation (3), $\gamma MSE(m) + c + e$ corresponds to PL, while $B\;\;b$ corresponds to V. One important difference between our formula and Hand and Posner's is that ours generalizes the discrete possibilities of Hand's formula (whether the speech causes harm or not) into the continuous event space measuring how much harm free expression imposes on others. Our continuous formula could therefore be used to achieve greater accuracy in consideration of the respective magnitudes of social benefits and harms from expression based on distorted information. We believe that the analysis provided in this article demonstrates the importance of specifying in detail which factors determine L and how the quality of private information among both expressers and receivers of expressed messages affects the criterion of social efficiency in cases of defamation, slander and libel.

Acknowledgements

We are deeply indebted to Kyu Ho Youm who currently holds the Jonathan Marshall First Amendment Chair at the University of Oregon for valuable discussions. We are also very grateful to the Editor Anders Nordgaard, an anonymous referee, Jack Knetsch and participants at the 11th annual conference of the Asian Law and Economics Association (AsLEA) held in Bangkok, Thailand and the monthly workshop of the Korean Institutional Economics Association held at SK Research Institute for helpful comments.

References

Arrow, K., 1997, Invaluable goods, *Journal of Economic Literature* **35**, 757–765.

Baker, C. E., 1989, *Human Liberty and Freedom of Speech*, Oxford University Press: New York, New York.

Berg, N., 2014, The consistency and ecological rationality schools of normative economics: Singular versus plural metrics for assessing bounded rationality, *Journal of Economic Methodology* **21**(4), 375–395.

Berg, N. and Gigerenzer, G., 2010, As-if behavioral economics: Neoclassical economics in disguise?, *History of Economic Ideas* **18**(1), 133–166

Beer, L. W., 1984, *Freedom of Expression in Japan: A Study in Comparative Law, Politics, and Society*, Kodansha International: Tokyo, Japan.

Dave, C. and Dodds, S., 2012, Nosy preferences, benevolence, and efficiency, *Southern Economic Journal* **78**(3), 878–894.

Dewey, J., 1925, *Experience and Nature*, Open Court: Chicago, IL.

Dworkin, R., 1977, *Taking Rights Seriously*, Harvard University Press: Cambridge, MA.

Glaeser, E. and Sunstein, C., 2014, Does more speech correct falsehoods?, *Journal of Legal Studies* **43**, 65–94.

[34] In this case, Hand affirmed convictions of 11 leaders of the Communist Party of the US for subversion under the 1940 Smith Act. Hand ruled that calls for the violent overthrow of the American government posed enough of a 'probable danger' to justify the restriction of free speech.

Hotelling, H., 1929, Stability in competition, *Economic Journal* **39**, 41–57.

Kim, J.-Y., 2006, Strict liability versus negligence when the injurer's activity involves positive externalities, *European Journal of Law and Economics* **22**, 95–104.

Mill, J. S., 1978, *On Liberty*, Hackett Publishing: Indianapolis, [Originally published in 1859].

Miller, A. and Perry, R., 2013, A group's a group, no matter how small: An economic analysis of defamation, *Washington and Lee Law Review* **70**, 2269–2336.

Pattanaik, P. K., 1994, On modeling individual rights: Some conceptual issues, In, edited by Arrow, Sen, and Suzumura. St. Martin's Press: New York.

Posner, R., 1986, Free speech in an economic perspective, *Suffolk University of Law Review* **20**, 1–54.

Radin, M. J., 1996, *Contested Commodities*, Harvard University Press: Cambridge, MA.

Sen, A., 1970, The impossibility of a Paretian liberal, *Journal of Political Economy* **78**, 152–157.

Sen, A., 1976, Liberty, unanimity and rights, *Economica* **43**, 217–245.

Sen, A., 1999, The possibility of social choice, *American Economic Review* **89**, 349–378.

Suzumura, K. and Xu, Y., 2001, Characterizations of consequentialism and nonconsequentialism, *Journal of Economic Theory* **101**, 423–436.

Suzumura, K. and Xu, Y., 2004, Welfarist-consequentialism, similarity of attitudes and Arrow's general impossibility theorem, *Social Choice and Welfare* **22**, 237–251.

Sunstein, C., 1993, *Democracy and the Problem of Free Speech*, The Free Press: New York, New York.

Sunstein, C., 2014, The Dark Side of the First Amendment, *Bloomberg View*, March 26.