APPENDIX[1]

Ordinal Status. Suppose there is a community of $n \in \mathbb{N}, n \geq 2$ individuals. The individuals are assumed to be identical in all but their strict rank ordering. The relative social position of individual *i* is given by r_i, with $r_i < r_j, \forall i < j$. The lowest ranking individual is thus noted by r_1 and the highest by r_n. Utility from status is given by the monotonically increasing utility function $U(.)$, with the utility of the individual with a status rank $r_i = U(r_i)$ and $U(0) = 0$.

Assume that the effect of defamation is to reduce one's social standing by $x > 0$ spots. It then follows that:

Corollary 1: With ordinal status, defamation is welfare neutral.

Proof:

Defamation moves the *t*th individual from position r_t to position $r_{t-x} < r_t$. But because rankings are relative, this implies that the individual previously in the r_{t-x} is moved up to the r_{t-x+1}position, and similarly for any individual above them with $r_i < r_t$. Thus, total utility, $\sum_1^n U(r_i)$, remains unchanged.■

Note that individuals are assumed here to have similar utility functions, but one might plausibly argue that individuals differ in the utility they draw from status. In such a case, defamation may lead to a net increase or decrease in utility, but there is no *a-priori* reason to assume any specific allocation of utility functions.

Cardinal Status. Suppose that individuals in the community are identical in all but their initial endowment of an intangible status good, with the total endowment being $S \in (0, \infty]$.[2] The endowment of status goods of the *i*th individual is denoted by s_i. This time, however, individuals do not care about their rank directly, but about their distance d_i from others. Individual *i*'s total distance from others is given by $d_i = \sum_{j=1}^{n}(s_i - s_j) = ns_i - S$.

Because status is defined as the accumulation of deference behavior or "pellets of peer recognition,"[3] we can think of defamation as destroying some of the target's 'pellets.' That is, defamation reduces the target's status goods by *x* units. Let d_t' be

---

[1] The model presented here addresses the standard view of defamation, which is decidedly ex-post in nature. The discussion in Part III addresses the ex-ante implications informally. There is limited, but growing, literature which studies formally the ex-ante effects of defamation law. *See generally*, Daniel Hemel, *Economic Perspectives on Free Speech*, *in* OXFORD HANDBOOK OF FREEDOM OF SPEECH (Frederick Schauer and Adrienne Stone eds., 2021); Oren Bar-Gill & Assaf Hamdani, *Optimal Liability for Libel*, 2 CONTRIBUTIONS TO ECON. ANALYSIS & POL'Y 1 (2003); Nuno Garoupa, *Dishonesty and Libel Law: The Economics of the "Chilling" Effect*, 155 J. OF INSTITUTIONAL & THEORETICAL ECONS. 284; Yonathan A. Arbel & Murat Mungan, *Regulating Speech with Bayesian Audiences* 15-22 (Univ. Ala. Legal Studs., Working Paper No. 3452662, 2020).

[2] Status goods can be thought of as the "accumulation of deference behavior", as in Michael Sauder, Freda Lynn & Joel M. Podolny, *Status: Insights from Organizational Sociology*, 38 ANNU. REV. SOCIOL. 267, 268 (2012).

[3] Robert K. Merton, *The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property*, 79 ISIS 606, 620 (1988).

the target's status ranking post defamation. We can now state the private harm to the target from defamation as:

$$|U(d_t)| - |U(d_t')| = |U(ns_t - \mathrm{S})| - |U\big(n(s_t - x) - (S - x)\big)|$$

That is, the target's original utility from her status goods less her utility from having *x* fewer units of the status good.[4] Note that, while defamation destroys some of the target's status goods, this loss is partially offset by the fact that there are fewer status goods to go around, which lowers the community average.

Finally, as we are considering the idea of risk to one's status, it will be fairly natural to assume that U'<0, U''>0.[5] We can now state the following proposition.

Proposition: Punching up: Defamation, on the margin, increases welfare if it is directed at a high-status individual. **Punching down:** Conversely, defamation reduces welfare on the margin if it is directed at a low-status individual.

Proof. Defamation destroys *x* units of status goods, and so the effect of defamation on total welfare W is:

$$W = \sum_{1}^{n-1} U(ns_i - S + x) + U(ns_t - nx - S + x)$$

The derivative with respect to *x* is:

$$\frac{dW}{dx} = \sum_{1}^{n-1} U'(ns_i - S + x) + (1 - n)U'(n(s_t - x) - S + x)$$

To understand the effect of small changes, evaluate at $x = 0$

$$\frac{dW}{dx}\Big|_{x=0} = \sum_{1}^{n-1} U'(ns_i - \mathrm{S}) + (1 - \mathrm{n})U'(ns_t - \mathrm{S})$$

Rearranging

$$\sum_{1}^{n} U'(ns_i - \mathrm{S}) - nU'(ns_t - \mathrm{S})$$

By concavity, the total marginal change in utility (the first expression on the left) will be larger than the change in the target's utility multiplied by *n* if and only if $s_t > \frac{S}{n}$. ■

Corollary 2: The optimal amount of defamation of an individual *t* is given by $x^* = \frac{ns_t - S}{n-1}$.

Proof. By the Proposition, defamation of any high-status individual is utility-maximizing on the margin, hence the optimal degree of defamation *x* is defined by $d_t = n(s_t - x) - S + x = 0$. Below that point, the individual becomes low-status, and punching-down is welfare-minimizing. Rearranging, this implies that $x^* = \frac{ns_t - S}{n-1}$. ■

[4] We consider the absolute value of the difference, as $U(r_t)$ and $U(r_t')$ may be negative.

[5] *See* GEOGGREY BRENNAN & PHILIP PETTIT, THE ECONOMY OF ESTEEM, 83-105 (2005). Another way to rationalize this assumption is to see that it is plausibly more consequential to a person's well-being to move close to those around them than to move farther apart. To the extent that wealth is also a status symbol, the marginal benefit from having *a* car rather than taking the bus is arguably greater than being able to afford a newer model.