True Myths

David K. Levine[1]

Abstract

Disagreement over social norms can lead to costly conflict. I use the word myth as a generic term for any type of narrative that communicates social norms. By communicating norms, myths can reduce disagreement and potentially improve welfare. To analyze this I study a simple model of public goods production in which the free rider problem is overcome by social norms supported by incentives in the form of external and internal punishments for failure to comply. In the context of competing social norms I consider "true" myths that support the first best. Do such a myths improve welfare? A true myth that is highly persuasive and pervasive leads to nearly first best welfare. To a surprising extent when either fails myths can be counterproductive.

Email address: david@dklevine.com (David K. Levine)
[1]Department of Economics, EUI and WUSTL
Acknowledgements: First version: September 10, 2022. I would like to thank Rohan Dutta, John Mair, Andrea Mattozzi and Salvatore Modica. I gratefully acknowledge support from the MIUR PRIN 2017 n. 2017H5KPLL_01.

1. Introduction

Disagreement over social norms can lead to costly conflict. I use the word myth as a generic term for any type of narrative that communicates social norms. By communicating norms, myths can reduce disagreement and potentially improve welfare. To analyze this I study a simple model of public goods production in which the free rider problem is overcome by social norms supported by incentives in the form of external and internal punishments for failure to comply. In the context of competing social norms I consider "true" myths that support the first best. Do such a myths improve welfare? Four main conclusions emerge from the analysis. The first is that a true myth that is highly persuasive and pervasive leads to nearly first best welfare. The second is that unpersuasive myths that simply make people feel bad without changing their behavior reduce welfare. Neither of these conclusions are terribly surprising. The third conclusion is that in a highly polarized environment only myths that are both highly persuasive and highly pervasive improve welfare. The fourth is that while ceasing to overproduce a public good is welfare improving, a myth that only marginally persuades people not to overproduce typically reduces welfare.

A great deal of economic research - from Marschak and Radner (1972)'s theory of teams to the recent literature on Bayesian persuasion stemming from Kamenica and Gentzkow (2011) - has examined the communication of facts: communications that may be either be true or false. A great deal of communication - all of fiction - has little to do with facts. One measure of the importance of facts versus myth is whether people listen to the news or to entertainment. A survey by Prior (2005) had randomly selected members of the US population rank their top four genres of television: only 16% indicated that news was either their top choice or second choice. Much entertainment contains information about social norms: even comic book heroes stand for "truth, justice and the American way." In general a good story often involves heroes whose social norms we should strive to emulate and villains whose should be punished for violating social norms. The myth communicated by the Harry Potter series is that a "good" person like Harry Potter is willing to risk everything to foil evildoers. I pick this example because J. K. Rowling, the author, is a compelling story teller. A compelling story is surely more likely to influence us than a boring or uninteresting story. The latter compels little attention, and soon is forgotten, while a compelling hero is someone we want to emulate. Take Paul Krugman as an example. He has written that his desire to become an economist was motivated by the fictional character of Hari Seldon in Isaac Asimov's Foundation trilogy. Like Harry Potter, *The Foundation* is pure fiction: nonetheless Hari Seldon's fictional theory of psycho-history was sufficiently compelling that it apparently led to at least one Nobel prize in economics.

It may well be that true stories are more compelling, all other things equal, than fictional ones. But I am fairly confident that dry recitations of facts - something we economists are justly famous for - are not particularly compelling. I doubt that an academic paper establishing that the social cost of carbon is $40 per ton is as likely to change social norms regarding carbon mitigation as an epic

film of environmental catastrophe, although the former is probably true and the latter probably not. As social norms determine not only our personal behavior, but also our political behavior, studying the role of myth and narrative seems important.

My plan is to introduce mythical communications, neither necessarily true or false, but communicating social norms, into an economic model. For concreteness I take the simplest setting in which social norms are important - the production of a public good. Here a social norm defines how much "should" be contributed to the public good. The importance of social norms in overcoming the free-rider problem is well established, beginning with the work of Coase (1960), and continuing on through Ostrom (1990), and many others. As social norms generally need to be enforced by punishing those who fail to comply, disagreement over social norms, that is, different people following different norms, can lead to socially costly punishment. Communication about what the social norm should be may alleviate this problem by reducing disagreement. My goal is to establish when this is true.

In the setting of public goods production I study three competing social norms: one for low contributions, one for middle contributions and one for high contributions. Of course a myth that supports an inefficient social norm is likely to reduce welfare, so I study the case where the middle and average contributions are first best and analyze a myth that supports the efficient level of contribution. Such a myth is "true" in the sense that if everyone adopted the social norm the first best would be attained.

To sustain a social norm for contributing to a public good in the face of the free rider problem, as indicated, social incentives are needed. These take two forms: external and internal. External incentives come about because people punish those who produce less than they do. Internal incentives come about because of the guilt that people feel for failing to live up to their own standards. I model myth as changing what those standards are, that is, they change internal incentives.

While the type of myths I study would lead to the first best if the social norms they promote were adopted by everyone, myths differ both in their persuasiveness - how much they lead people to change their standards, and their pervasiveness - how many people are influenced by the myth. In this setting I show how and when a myth that fails to be both highly persuasive and pervasive can be counterproductive, lowering rather than increasing welfare over the status quo. This is the case when polarization is high. Myths that have little persuasiveness are always counterproductive. With a low degree of polarization moderately persuasive myths that are also pervasive give large welfare increases while more persuasive myths can be counterproductive regardless of how pervasive they are. The welfare function is discontinous at points where extreme types change their behavior, and persuasiveness just below the lower switchpoint or just above the higher switchpoint are especially bad.

Related Work

The idea that myth is functional in communicating social norms is widespread in the literature on sociology and anthropology and is often associated with the work of Durkheim - see for example Cohen (1969). In economics, Hickson and Thompson (1991) developed the idea that we emulate our heroes. I have not, however, been able to find an economics literature on myth - a Google Scholar search for "economics myth" yields many hits for papers indicating that economics itself is a myth, but not for the economic study of myth.

The functional role of myth in the form of narrative is closely connected to the role of narrative in establishing identity. This idea was developed experimentally in Tajfel et al (1971) and is now used in many experimental economic studies. This literature confirms the idea that narrative impacts on identity. Work by Ely, Frankel and Kamenica (2015) and Corrao, Fudenberg and Levine (2022) examine the role of surprise in writing a compelling narrative.

The model of social norms that underlies the theory here derives from the literature on internalization of social norms and on the use of punishment in enforcing social mechanisms. The model of internalization appears in the literature on ethical voters, such as Feddersen and Sandroni (2006) and Coate and Conlin (2004), and the literature on the warm glow effect, such as Andreoni (1990) and Palfrey and Prisbrey (1997). The model of punishment used here was introduced by Levine and Modica (2016) and used to study voter turnout by Levine and Mattozzi (2020). In Dutta, Levine and Modica (2021) we combined the two ideas, and that paper is the starting point of the model here. In the earlier work on the equilibrium path punishment arose with a common social norm due to imperfect observability of contributions. Here on the equilibrium path punishment arises due to disagreement over the social norm, so I simplify by assuming that contributions are perfectly observed.

2. The Model

There are two periods $t = 1, 2$, before and after the arrival of a myth, and there is a group with a continuum of members i uniformly distributed over the unit interval. There is a stage game that takes place in each period and the myth that arrives between periods determines the stage game in the second period.

In the period t stage game each member i chooses how much effort $x_t^i \geq 0$ to contribute to a public good. Effort has constant unit marginal cost. It is used to produce a public good, the output of which depends upon the average effort $x_t = \int x_t^i di$ and yields a social benefit to all members of $f(x_t)$. Here $f(0) = 0$, and f is strictly increasing, strictly concave, and differentiable for $x_t > 0$. Base utility of a member in period t is then $f(x_t) - x_t^i$ and input is normalized with $f'(1) = 1$, so that the *first best* in each period is a unit of output. The function $f(x_t) = 2\sqrt{x_t}$ satisfies these properties and I will use it in examples. Intertemporal preferences are the average present value of the two periods with respect to some non-negative discount factor.

The effect of any individual member in a large group on average effort is negligible, so there is a severe free-rider problem. This free rider problem is solved using two types of incentives: external and internal. External incentives arise from punishment by others for failing to do one's share. Specifically, member i is punished by everyone who provides strictly more effort. The utility cost of the punishment is $P \geq 0$ so if $F_t(x)$ is the cdf of input at time t then absent internal incentives the utility of member i in period t is given as $f(x_t) - x_t^i - \left(1 - F(x_t^i)\right) P$. Notice that there is no disagreement over how much to punish: the model is designed to analyze disagreement over how much to produce.

Internal incentives for member i in period t are given by a *guilt function* $g_t^i(x_t^i) \geq 0$, weakly decreasing and right continuous. A *simple* guilt function has a quota y_t^i and guilt $g_t^i(x_t^i) = \gamma \geq 0$ for $x_t^i < y_t^i$ and $g_t^i(x_t^i) = 0$ for $x_t^i \geq y_t^i$. Here γ represents the disutility from failing to do one's share by contributing up to the quota y_t^i. Hence expected utility in period t is $f(x_t) - x_t^i - \left(1 - F(x_t^i)\right) P - g_t^i(x_t^i)$ and averaged over individuals this defines *welfare*.

In between the two periods a myth arrives. The myth supports a particular guilt function $h(x)$ and is characterized by its pervasiveness $0 \leq \nu \leq 1$ and its persuasiveness $0 \leq \sigma \leq 1$. Pervasiveness is the fraction of randomly chosen group members who hear the myth and are influenced by it and persuasiveness is how much their guilt function is affected by the myth. Specifically, a group member who hears the myth and is *influenced* has a second period guilt function $g_2^i(x_2^i) = (1 - \sigma) g_1^i(x_2^i) + \sigma h(x_2^i)$. In particular, if the myth is completely persuasive with $\sigma = 1$ then the mythical guilt function $h(x)$ is adopted, while if the myth is completely unpersuasive with $\sigma = 0$ then the second period guilt function is the same as in the first period. Group members who do not hear or pay attention to the myth so are *uninfluenced* are unpersuaded, and their second period guilt function is the same as the first, $g_2^i(x_2^i) = g_1^i(x_2^i)$. Note that I distinguish being influenced from being informed: it possible to hear a myth but find it boring or irrelevant.

Initially in period 1 there are three types of group members $\tau \in \{L, M, H\}$ with simple guilt functions having quotas y_τ. The middle types M have a quota equal to the first best, $y_M = 1$, the low types L have a quota $y_L = 1 - \Delta$ and the high types H have a quota $y_H = 1 + \Delta$ where $0 < \Delta < 1$. There are equal fractions $0 < \phi < 1/2$ of the low and high types. This means that if all types produce their quota output is the first best. Never-the-less inefficiency can arise because the disagreement over the quotas leads to punishment.

The myth supports the middle quota: $h(x) = g_1^M(x)$. This means that the middle types are not affected by the myth. The lower and upper types are split into two sub-types, for each type there are $(1 - \nu)\phi$ uninfluenced whose guilt functions are not changed, and $\nu\phi$ influenced. For the low types the second period guilt function for the influenced is $g_2^L(x_2^i) = \gamma$ for $x_2^i < y_L$, $g_2^L(x_2^i) = \sigma\gamma$ for $y_L \leq x_2^i < y_M$, and $g_2^L(x_2^i) = 0$ for $x_2^i \geq y_M$. For the high types the second period guilt function for the influenced is $g_2^H(x_2^i) = \gamma$ for $x_2^i < y_M$, $g_2^H(x_2^i) = (1 - \sigma)\gamma$ for $y_M \leq x_2^i < y_H$, and $g_2^H(x_2^i) = 0$ for $x_2^i \geq y_H$.

Because group members are negligible our notion of equilibrium is that of

open-loop equilibrium in which the second period equilibrium does not depend upon the that first period choice of any individual group member. This means that intertemporal preferences are irrelevant and that each member behaves as if myopic. These open-loop equilibria are shown by Fudenberg and Levine (1988) to be approximate subgame perfect equilibria of underlying finite player games. Within the the open-loop equilibria I restrict attention to those in which all group members with the same guilt function contribute the same amount.

I want to study a situation where initially there are three meaningful social norms corresponding to the three types, and each type finds it optimal to implement their own social norm. This means that the status quo, in the form of the first period quotas, plays a special role. Hence I will restrict attention to equilibria in which the status quo is *respected*. For those group members who have a simple guilt function I take this to mean that their contribution is equal to their quota. In the first period this completely determines a unique equilibrium outcome in which each type contributes their respective quota. Such an equilibrium does not exist for all parameter values, and I will study only those parameter values for which one does exists. I call the parameters γ, Δ *feasible* if for any distribution over types there exists a P for which the status quo is respected in the first period and say that such a P supports the status quo. The parameters are *non-trivia*l if any such supporting P is strictly positive. In the Appendix Theorem 4.1 it is shown that the parameters γ, Δ are feasible if and only if $\gamma \geq 1$ and $\Delta \leq \gamma/2$ and P is supporting if and only if $\Delta \geq P \geq \Delta - (\gamma - 1)$. An immediate implication is that feasible parameters are non-trivial if and only if $\gamma - 1 < \Delta$. I will analyze only feasible parameters with a supporting punishment.

For influenced second period group members I say that the status quo is respected if when their equilibrium contribution differs from the status quo it is strictly optimal with respect to the status quo beliefs that all group members contribute their first period quotas. In other words, nobody switches from the status quo unless it is strictly optimal to do so.

3. Myth and Welfare

My goal is to assess when a myth increases and decreases welfare. Essential to this is the fact that there are three types of second period equilibrium separated by switchpoint values of persuasiveness. These switchpoints correspond to the indifference of the low types to switching, given by

$$\underline{\sigma} = \frac{\Delta - (1 - 2\phi)P}{\gamma}$$

and to the indifference of the high types to switching, given by

$$\overline{\sigma} = \frac{\gamma + \phi P - \Delta}{\gamma}.$$

This is shown in Theorem 4.2 in the Appendix which characterizes equilibrium and establishes the main result of this paper:

Theorem 3.1. *For any feasible parameters γ, Δ with a supporting P we have $0 < \underline{\sigma} < \overline{\sigma} < 1$ and there is a unique status quo respecting equilibrium given by*

In the first period each type contributes their quota.

In the second period there are three types of equilibrium: weak, moderate, and strong.

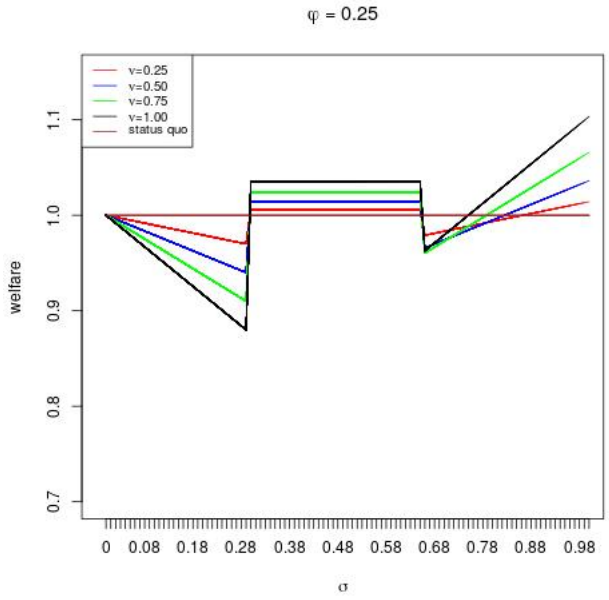
Weak equilibrium *occurs for $\sigma \leq \underline{\sigma}$ and every type contributes their first period quota. Welfare is linear and decreasing in σ and linear and decreasing in ν.*

Moderate equilibrium *occurs for $\underline{\sigma} < \sigma \leq \overline{\sigma}$. Influenced low types contribute y_M and all other types contribute their first period quota: output is greater than the first best level. Welfare is constant in σ.*

Strong equilibrium *occurs for $\sigma > \overline{\sigma}$. Influenced types contribute y_M and all other types contribute their first period quota: output is equal to the first best level. Welfare is linear and increasing in σ.*

Welfare jumps at the switchpoints $\underline{\sigma}, \overline{\sigma}$.

The theorem is well illustrated by a numerical example. The production function is $f(x_t) = 2\sqrt{x_t}$. Parameters are chosen to be feasible: $\gamma = 1.5 > 1$, $\Delta = 0.6 < 1, \gamma/2$ and $\Delta > \gamma - 1 = 0.5$. The punishment $P = 0.3$ is chosen to satisfy $0.6 = \Delta > P > \Delta - (\gamma - 1) = 0.1$. The first set of examples plots the ratio of welfare in the second period to that in the first when $\phi = 0.25$ for different values of σ, ν.



The graph illustrates the theoretical result that low levels of persuasiveness are counterproductive until the moderate equilibrium is reached. The moderate equilibria are pretty good from a welfare point of view: increasing persuasiveness beyond the moderate range drops welfare substantially, and indeed below the status quo, and it requires substantial persuasiveness before the welfare level of the moderate range is reached again. Crucial to the implications of the theory are the substantial welfare jumps at the switchpoints.

The idea of the proof of Theorem 3.1 can be understood by tracing out what happens in the graph as persuasiveness σ increases from zero to one. Persuasiveness has no effect on the guilt of the middle types, so they never move. Uninfluenced types never move. Initially the myth is totally unpersuasive so the influenced extreme types also remain at the status quo. As the persuasiveness of the myth increases this makes the influenced low types feel increasingly guilty, but they still strictly prefer to contribute y_L so this just lowers their utility. The influenced high types do not feel guilty, but the guilt they feel from switching to y_M is reduced. Still they strictly prefer to contribute y_H. The sole initial effect of increasing persuasiveness is to make the influenced low types feel guiltier, so overall welfare is reduced. Increasing pervasiveness further reduces welfare by increasing the number of low types who are influenced and feel guilty.

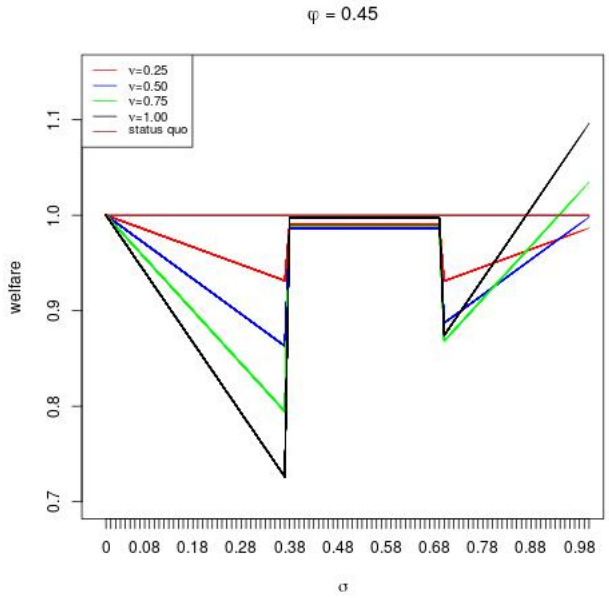
The fact that $0 < \underline{\sigma} < \overline{\sigma} < 1$ means that the influenced low types become indifferent first. As σ increases past the point of indifference, the strict best response is for the influenced low types to switch. Other low types switching increases the incentives of a low type to switch and has no effect on the incentives of the middle or high type, so this results in a status quo respecting equilibrium. As the influenced low types are indifferent to switching at $\underline{\sigma}$ welfare changes only due to the two effects that they do not internalize: the increased output of the public good that benefits everyone and the fact that the influenced low types now punish the uninfluenced low types. As we shall see, the output effect generally dominates the punishment effect, so that the jump in welfare will typically be upwards.

At this moderate equilibrium nobody feel guilty so increasing persuasiveness no longer makes a difference for welfare until $\overline{\sigma}$ is reached. At this point the influenced high types become indifferent. As persuasiveness increases the influenced high types strictly prefer to switch. Their switching increases the incentive of both the low and high types to switch, but the influenced low types have already switched, so this is indeed a status quo respecting equilibrium. Output jumps back down to the first best level reducing welfare, but this is offset by reduced punishment of the middle types. Again the output effect generally dominates the punishment effect so the jump is typically down. Theorem 4.2 shows that the jump up is greater than the jump down exactly when $\nu > (3\phi - 1)/(2\phi)$. This always holds when $\phi \leq 1/3$ as in the example. Further increases in persuasiveness have no effect on behavior, but reduces the guilt felt by the influenced high types for switching, so increases welfare.

Polarization

The qualitative and quantitative features of the impact of a myth on welfare depends crucially on the degree of polarization. If $\phi = 1/3$ there are initially equal numbers of each type. If $\phi < 1/3$ then there are initally more middle types than any other type: this means that there is not so much polarization, and we refer to this as *low polarization*. This is the case in the first example, where $\phi = 0.25$. If $\phi > 1/3$ there there are initially less middle types than any other type, and we refer to this as *high polarization*.

While it is natural to think that increasing the number of middle types, that is, reducing polarization, decreases conflict, this is only true when polarization is low. In the status quo there are $1 - \phi$ middle and high types punishing the ϕ low types and ϕ high types punishing the $1 - 2\phi$ middle types so the total amount of punishment is $(2\phi - 3\phi^2)P$. Hence the expected amount of punishment is increasing in ϕ, that is to say decreases as polarization is reduced, exactly when there is low polarization. This suggests that myth is more likely to be counterproductive when polarization is high, and this is illustrated in the second numerical example in which ϕ is increased from 0.25 to 0.45. Again, welfare in the second period is plotted relative to than in the first for different values of σ, ν.



In comparison to $\phi = 0.25$ we see that the welfare loss due to guilt increases much more quickly as persuasiveness increases as there are more low types to feel guilty. More significant is the fact that the moderate equilibrium is worse than the status quo, while it was substantially better when $\phi = 0.25$. For lower values of $\nu \in \{0.25, 0.50\}$ even full persuasiveness is not enough to lift welfare to the

status quo level. Both high persuasiveness and high pervasiveness are needed for a welfare improvement when polarization is high. It is worth noting as well that when polarization and pervasiveness are high the loss from low persuasiveness, about 30%, is considerably higher than the gain from high persuasiveness, which is only about 10%.

The Output and Punishment Effects

The welfare jumps at the switchpoints consist of two offsetting effects. Output changes by $\nu\phi\Delta$ due to $\nu\phi$ extreme types switching to y_M. This gives rise to a welfare change of $f(1+\nu\phi\Delta) - f(1)$.

At the lower switchpoint the benefit of increased output is offset by the fact that $\nu\phi$ influenced low types have now switched to y_M and begin punishing the $(1-\nu)\phi$ uninfluenced low types. This decreases welfare by $\nu\phi(1-\nu)\phi P$. At the upper switchpoint the loss from decreased output is offset by the fact that $\nu\phi$ influenced high types have now switched to y_M and no longer punish the $(1-2\phi+\nu\phi)$ members who were contributing y_M. This increases welfare by $\nu\phi(1-2\phi+\nu\phi)P$. In both cases as feasibility requires $\Delta > P$ the punishment effect is strictly less than $\nu\phi\Delta$.

How large the output effect is depends on the production function. Since it is assumed differentiable it cannot be completely flat to the right of $x_t = 1$, but it can be arbitrarily close, meaning that the production effect can be arbitrarily small. On the other hand, the slope of the production function at $x_t = 1$ is one and the function is concave, so $f(1+\nu\phi\Delta) - f(1) < \nu\phi\Delta$. However, as $\nu\phi\Delta$ is small, for reasonably well behaved production funtions the linear approximation at $x_t = 1$ is a good one so that in fact $f(1+\nu\phi\Delta) - f(1) \approx \nu\phi\Delta$. With this approximation the production effect always dominates the punishment effect, which is why in the numerical examples the welfare function first jumps up then jumps down.

4. Conclusion

I can illustrate the theory with reference to the climate debate. Here the public good is reducing carbon usage. We can think of low types as climate change deniers whose social norm is to advocate buying gas guzzling trucks, flying on private jets, and gobbling steaks and burgers. The high types we may think of as green advocates whose social norm is to advocate being vegetarian, and travelling only by train or boat. The middle types are economists whose social norm is to advocate a substantial but not excessive carbon tax. The narrative of the low types is that the story of climate change is a false narrative driven by evil rent-seekers trying to promote their own "green" businesses. That of the high types is that evil industrialists are willing to destroy civilization if it means they can afford a few more super yachts. These are great and compelling stories, while the dry recitation of facts favored by economists are not.

Trusting that the economists are right, the myth of a substantial but not excessive carbon tax is a "true" myth. The theory indicates that it is a welfare

improving myth only if it is pervasive enough to matter, and persuasive enough at least to persuade the low types to switch. Here I would channel the work of Hassler, Krusell and Olovsson (2018) on downside and upside risk to suggest how a skilled story-teller might construct a compelling narrative: the disaster scenarios of the high types make an interesting, exciting and memorable story. As the low types point out, some of this narrative is driven by self-seeking rent-seekers and these disasters are not so likely to happen. Never-the-less, it is surely a good idea to take low cost steps to reduce the chances that they do happen? I wonder, however, given the discontinuity at the upper switchpoint if we really want to try to convince the high types to switch? Perhaps it would be safer just to be persuasive enough to get the low types to switch?

I want to wrap this up by discussing pervasiveness. If a myth is not very pervasive it is has little effect on welfare and so does not matter much. The pervasiveness of a myth depends upon how many people are influenced by it - how many people hear the myth and pay attention to it. In practice the exposure a myth receives depends on word-of-mouth. If we all tell our friends about a good article, book, or movie, it becomes "viral" and pervasive. This means that in our own behavior we can influence the pervasiveness of myths by choosing what to tell our friends, and what to dissuade our friends from telling others. The results here provide some guidance from a welfare point of view about which "true" myths should be encouraged and which discouraged.

References

Andreoni, J. (1990): "Impure altruism and donations to public goods: A theory of warm-glow giving," *Economic Journal* 100: 464-477.

Asimov, Isaac (1951): *Foundation*, Gnome Press.

Coase, R. H. (1960): "The Problem of Social Cost," *Journal of Law and Economics* 3: 1-44.

Coate, S., M. Conlin (2004): "A Group Rule–Utilitarian Aapproach to Voter Turnout: Theory and Evidence," *American Economic Review* 94: 1476–1504.

Cohen, P. S. (1969): "Theories of myth," *Man* 4: 337-353.

Corrao, Roberto, Drew Fudenberg and David K. Levine (2022): "Adversarial forecasters, surprises and randomization," mimeo EUI.

Dutta, R., D. K. Levine and S. Modica (2021): "The Whip and the Bible: Punishment Versus Internalization," *Journal of Public Economic Theory*, 23: 858-894

Ely, J., A. Frankel, and E. Kamenica (2015): "Suspense and surprise," *Journal of Political Economy* 123: 215–260.

Feddersen, T., A. Sandroni (2006): "A Theory of Participation in Elections," *American Economic Review* 96: 1271–1282.

Fudenberg, D. and D. K. Levine (1988): "Open and Closed-Loop Equilibria in Dynamic Games With Many Players," *Journal of Economic Theory*, 44: 1-18

Hickson, C. R. and E. A. Thompson (1991): "A new theory of guilds and European economic development," Explorations in Economic History 28: 127-168.

Kamenica, E. and M. Gentzkow, M. (2011): "Bayesian persuasion," *American Economic Review* 101: 2590-2615.

Krugman, Paul (2012): "Paul Krugman: Asimov's Foundation novels grounded my economics," *The Guardian*, December 4.

Hassler, J., Krusell, P., and Olovsson, C. (2018): "The consequences of uncertainty: climate sensitivity and economic sensitivity to the climate," *Annual Review of Economics* 10: 189-205.

Levine, David and Salvatore Modica (2016): "Peer Discipline and Incentives within Groups", *Journal of Economic Behavior and Organization* 123: 19-30

Levine, David K. and Andrea Mattozzi (2020): "Voter Turnout with Peer Punishment," *American Economic Review* 110: 3298–3314.

Marschak, Jacob and Roy Radner (1972): *Economic Theory of Teams,* Cowles Commission.

Ostrom, Elinor (1990): *Governing the commons: The evolution of institutions for collective action*, Cambridge university press.

Palfrey, T. R. and Prisbrey, J. E. (1997): "Anomalous behavior in public goods experiments: How much and why?" *American Economic Review*, 829-846.

Prior, Markus (2005): "News vs. Entertainment: How Increasing Media Choice Widens Gaps in Political Knowledge and Turnout," *American Journal of Political Science* 49: 577–92.

Rowling, J. K. (1997): *Harry Potter and the Philosopher's Stone*, Bloomsbury.

Tangney, J. P., J. Stuewig and D.J. Mashek (2007): "Moral emotions and moral behavior," *Annual Review of Psychology* 58: 345-372.

Tangney, J. P. and R. L. Dearing, R. L. (2003): *Shame and Guilt.* Guilford Press.

Tajfel, H., M.G. Billig, R.P. Bundy, and C. Flament (1971): "Social categorization in intergroup behavior," *European Journal of Social Psychology* 1: 149-178.

Townsend, R. M. (1994): "Risk and insurance in village India," *Econometrica*, 539-591.

Appendix

Theorem 4.1. *The parameters γ, Δ are feasible if and only if $\gamma \geq 1$ and $\Delta \leq \gamma/2$ and P is supporting if and only if $\Delta \geq P \geq \Delta - (\gamma - 1)$.*

Proof. With a simple guilt function there must be an optimum at $0, y_L, y_M$ or y_H. Since $\Delta < 1$ zero contribution is ruled out if and only if the high type prefers not to deviate: since deviating to zero results in punishment by all types this is $P + \gamma \geq y_H = 1 + \Delta$. It must also be the case that the high type prefers not to deviate to $y_L$or y_M. If the entire population consists of low types deviating to y_L invokes no punishment, has a guilt cost of γ and gains 2Δ in cost reduction, so the condition is $2\Delta \leq \gamma$. On the other hand, the low types must prefer not to deviate to y_M or y_H. If there are only middle types there is a punishment reduction of P for providing y_M and a cost increase of Δ, so the condition is $\Delta \geq P$.

We see then that feasibility requires that $\Delta \geq P \geq 1 + \Delta - \gamma$ and this is possible if and only if $\gamma \geq 1$. Suppose on the other hand that $\gamma \geq 1$, $\Delta \leq \gamma/2$ and $\Delta \geq P \geq 1 + \Delta - \gamma$. Since $\Delta \leq \gamma/2$ the middle and high types prefer their own quota to any lower quota since the greatest gain is 2Δ and the loss is at least γ. All types prefer their own quota to producing zero since the loss is $P + \gamma$ while the greatest possible gain is $1 + \Delta$. Producing above quota reduces punishment by at most P while incuring a cost of at least Δ so neither the low or middle type wish to do so. Hence $\gamma \geq 1$ and $\Delta \leq \gamma/2$ are both necessary and sufficient as asserted.

Finally, suppose it is possible to choose $P = 0$. Then $\gamma \geq 1 + \Delta$, and conversely if $\gamma \geq 1 + \Delta$ then it is possible to choose $P = 0$. Hence the condition for non-triviality is $\gamma < 1 + \Delta$. As $\Delta \leq \gamma/2$ it must be that $\gamma < 2$. □

Theorem 4.2. *For any feasible parameters γ, Δ with a supporting P we have $0 < \underline{\sigma} < \overline{\sigma} < 1$ and there is a unique status quo respecting equilibrium given by*

In the first period each type contributes their quota. Welfare is $W_1 = f(1) - 1 - \phi(2 - 3\phi)P$.

In the second period there are three types of equilibrium: weak, moderate, and strong.

Weak equilibrium occurs for $\sigma \leq \underline{\sigma}$ and every type contributes their first period quota. Welfare is $W_w = f(1) - 1 - \phi(2 - 3\phi)P - \sigma\nu\phi\gamma$.

Moderate equilibrium occurs for $\underline{\sigma} < \sigma \leq \overline{\sigma}$. Influenced low types contribute y_M and all other types contribute their first period quota: output is $y_m = y_M + \phi\Delta\nu$, and welfare is $W_m = f(y_m) - y_m - \phi\left((1-\nu)(1-(1-\nu)\phi) + (1 - 2\phi + \nu\phi)\right)P$.

Strong equilibrium occurs for $\sigma > \overline{\sigma}$. Influenced types contribute y_M and all other types contribute their first period quota. $W_s = f(1) - 1 - (1-\nu)\phi(2 - 3(1-\nu)\phi))P - (1-\sigma)\nu\phi\gamma$.

Welfare satisfies $W_w(\underline{\sigma}) < \lim_{\sigma \downarrow \overline{\sigma}} W_s(\sigma)$ exactly when

$$\nu > \frac{3\phi - 1}{2\phi}.$$

Proof. The first period equilibrium is unique and status quo by Theorem 4.1. There is no guilt, output is the first best by construction, so welfare is $f(1)-1$ minus the cost of punishment. The low type is punished by the middle and high type, so receives an punishment of $(1-\phi)P$. The middle type is punished by the high type so receives an expected punishment of ϕP. The high type is not punished. As there are ϕ low types and $1-2\phi$ middle types the average expected punishment is

$$\phi(1-\phi)P + (1-2\phi)\phi P = \phi(2-3\phi)P$$

giving welfare as indicated.

In the second period the middle type and uniformed types optimally contribute their status quo amounts. The influenced low and high types may switch to contributing the middle quota $y_M = 1$. By definition of respecting the status quo if it is weakly optimal for an influenced type to play the status quo against all members playing the status quo they cannot switch. I first determine when it is weakly optimal not to switch.

The high type gains Δ by switching to middle. Punishment goes from zero to ϕP due to punishment by the other high types. The guilt of switching is $(1-\sigma)\gamma$. Hence the high type weakly prefers not to switch when $\Delta \leq \phi P + (1-\sigma)\gamma$. Hence $\overline{\sigma}$ is the cutoff for indifference. Since $\gamma \geq 1$ and $\Delta < 1$ we have $\overline{\sigma} > 0$. Since Theorem 4.1 requires that $\Delta \geq P > \phi P$ we have $\overline{\sigma} < 1$. Hence the high type strictly prefers not to switch if and only if $\sigma \leq \overline{\sigma}$.

The low type loses Δ by switching to middle. Punishment is reduced by $(1-2\phi)P$ as punishment from the middle types is escaped. In addition the guilt from sticking to the status quo is reduced from $\sigma\gamma$ to zero. Hence the low type weakly prefers not to switch when $\Delta \geq (1-2\phi)P + \sigma\gamma$. Hence $\underline{\sigma}$ is the cutoff for indifference. Since Theorem 4.1 requires that $\Delta \geq P > (1-2\phi)P$ we have $\underline{\sigma} > 0$. since $\Delta < 1$ and $\gamma \geq 1$ we have $\underline{\sigma} < 1$. Hence the low type weakly prefers not to switch if and only if $\sigma \leq \underline{\sigma}$.

We next show that $\overline{\sigma} > \underline{\sigma}$. From Theorem 4.1 we have $\gamma \geq 2\Delta$. Hence $\gamma > 2\Delta - (1-\phi)P$. This can be rearranged as $\gamma + \phi P - \Delta > \Delta - (1-2\phi)P$, that is, the numerator of $\overline{\sigma}$ is strictly larger than that of $\underline{\sigma}$, so indeed $\overline{\sigma} > \underline{\sigma}$.

Weak equilibrium now follows as both types weakly prefer not to switch, so the status quo itself is the unique status quo respecting equilibrium. Welfare is that of the first period status quo equilibrium minus the additional guilt felt by the low types: this is $\sigma\gamma$ and is felt by the $\nu\phi$ influenced low types, giving welfare as indicated.

Next observe that the incentive to switch depends on the choices of others only through punishment: the contribution gain or loss, and guilt gain or loss does not depend upon what others are doing. Hence low types switching has no effect on the incentives of the high type to switch: the punishment from switching depends only on the number of high types. Low types switching increases the incentive of other low types to switch since it increases the punishment for not switching. High types switching increases the incentive for low types to switch as it reduces the punishment after switching, and increases the incen-

tives for high types to switch, since it reduces the punishment for switching. This gives the moderate and strong cases. In the moderate case the influenced high types continue to find it optimal not to switch - what the low types do does not matter. On the other hand, when all the influenced low types strictly prefer to switch at the status quo, they must do so, and when they do so they find it strictly optimal since the incentive to switch is increased. In the strong case the influenced low types have already switched, and the influenced high types switching just makes this better, while the previous argument applies to the high types. This proves that in each case there is a unique status quo respecting equilibrium as described.

It remain to find welfare in moderate and strong cases.

For the moderate case output is increased because ν of the low types increase their contributions by Δ. As there are $\nu\Delta$ of them this increases average contributions by $\nu\phi\Delta$ as asserted. Nobody feels guilty in this equilibrium, so it remains to work out the punishment cost. After switching there are $(1-\nu)\phi$ low types and $(1-2\phi+\nu\phi)$ middle types. The former are punished all other types, so get an expected punishment of $(1-(1-\nu)\phi)P$. The latter are punished by the high types so get an expected punishment of ϕP. Averaging with the number of types of each gives welfare as indicated.

For the strong case output is again first best as equal numbers of low and high types switch to middle. Punishment cost is determined by replacing ϕ with $\nu\phi$ in the status quo punishment cost. There is also a guilt cost from high types who switch of $(1-\sigma)\gamma$. Averaging with the number of high types gives welfare as indicated.

Finally, we prove that $W_w(\underline{\sigma}) < \lim_{\sigma\downarrow\overline{\sigma}} W_s(\sigma)$. The output effect is a wash, so this is just a comparison of the punishment effects. At the bottom $\nu\phi$ influenced low types start to punish $(1-\nu)\phi$ uninfluenced low types, while at the top $\nu\phi$ influenced high types stop punishing the $(1-2\phi+\nu\phi)$ types who are contributing y_M. The condition $1-2\phi+\nu\phi > \phi-\nu\phi$ is equivalent to that in the Theorem. □