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? I study both the short-run and the long-run. 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 in the short-run. Hence true myths may be a costly short-run investment in a good long-run outcome.

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? I study both the short-run and the long-run. Four main conclusions emerge from the short-run 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. By contrast, in the long-run true myths lead ultimately to the first best. Hence true myths may be a costly short-run investment in a good long-run outcome.

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 in Rowling (1997) is that a "good" person like Harry Potter is willing to risk everything to foil evil-doers. 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. In Krugman (2012) he wrote 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, Asimov (1951) 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 examine the case where the middle contributions are first best and analyze myths that supports that efficient level of contribution. Such myths is "true" in the sense that if everyone adopted the promoted 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 in the short-run 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. The

short-run landscape is a treacherous one: a slightly more persuasive myth may result in large welfare gains or losses. By contrast I show that when myths repeatedly arrive eventually they put an end to conflict and result in the first best.

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). That literature argues also that myth is important in managing conflict: see for example Brenneis (1988). In economics 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. In a similar vein Shiller (2017) and Shiller (2019) establishes that there is little economic research on narrative although there is a great deal of research in the other social sciences. Somewhat ironically, his focus is then on the idea that economics itself is a myth. I have not been able to find an economics literature in which myth or narrative can reduce conflict or, indeed, serve any other purpose.

The functional role of myth in the form of narrative is closely connected to the role of narrative in establishing identity. Hickson and Thompson (1991) developed the idea that we emulate our heroes. The role of narrative in establishing identity was developed experimentally in Tajfel et al (1971) and is now used in many experimental economic studies. Similarly Michalopoulos and Xue (2021) shows that across a broad range of societies social norms reflect social myths.

Concerning what constitutes an effective myth, work by Ely, Frankel and Kamenica (2015) and Corrao, Fudenberg and Levine (2022) examine the role of surprise in writing a compelling narrative. Concerning the impact of narrative Benabou, Falk and Tirole (2020) study how it changes pyschological beliefs about the tradeoff between private benefits and social costs. Eliaz and Spiegler (2020) study the impact on beliefs about causality. All of this work, however, considers purely psychological and informational effects of narrative and not its role in communicating social norms. Myth simply spreads truth or lies and there is no costly conflict over social norms and hence no role for myth in reducing social conflict.

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. Subsequently I will study additional periods in which myths arrive.

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* and 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 $0 < \Delta < 1$. The fraction of type τ is $\phi_\tau > 0$.

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 τ there are $(1-\nu)\phi_\tau$ uninfluenced whose guilt functions are not changed, and $\nu\phi_\tau$ 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 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 the first period *status quo* equilibrium in which these quotas are adopted. As the quotas are not an equilibrium for all parameter values I will study only those parameter values for they are. 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 3 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.

Status quo output is either inefficiently low if $\phi_L > \phi_H$ or high if $\phi_H > \phi_L$. The *balanced case* in there are equal fractions $\phi_L = \phi_H = \phi$ of the high and low types is a useful benchmark for isolating the role of myth in reducing conflict. In this case in the first period when if all types produce their quota output is the first best. Never-the-less the status quo is inefficient because disagreement over the quotas leads to punishment.

3. The Shadow of the Past

My basic notion of equilibrium is open-loop Nash equilibrium, but I also want to capture the idea that the first period status quo has meaning in the second period. For example, if the status quo is an equilibrium in the second period I would not expect people to spontaneously switch to another equilibrium. In the current context of social norms there is an extensive empirical literature establishing that people do not do this. Acemoglu and Robinson (2001) give evidence of persistence of the status quo on the order of four centuries. Bigoni et al (2013) have evidence of a similar effect in Italy over nearly nine centuries and Belloc, Drago and Galbiati (2016) point to persistence In Italy that also lasts centuries. Dell and Querubin (2018) have highly persuasive evidence for persistence in Vietnam on the order of a century and a half.

A brief discussion of Dell and Querubin (2018) makes the point. Here neighboring villages have different social norms dating back to a time when they were ruled by different empires. Despite the fact that the poorer villages could adapt the norms of the more prosperous villages they do not do so. Notice that the issue here is not individual behavior: no individual in the poorer village would benefit from adapting the norm of a more prosperous village, but if everyone were to do so all would benefit.

To capture the idea that people stick with the status quo I introduce a refinement of Nash equilibrium I call *status quo respecting*. It is grounded in recent work on learning in games which seeks to explain why we see Nash equilibrium in the first place. This literature, beginning with Foster and Young (2003) and including Foster and Young (2006), Young (2009) and Block, Fudenberg and Levine (2019) argues that if people use stochastic learning rules that are conservative in the sense that they tend to stick with the status quo then we will observe Nash equilibrium most of the time.

I base my discussion of status quo respecting equilibrium on Block, Fudenberg and Levine (2019) who distinguish between content and discontent players. Content players continue to play as they have been. A content player who is not playing a best response eventually becomes discontent. Discontent players know that they can do better and strive to find out how. If the status quo is a Nash equilibrium then everyone is content and that is how players play: this captures the simple idea that if the status quo is an equilibrium it remains an equilibrium.

I would also like to say what happens if some players become discontent: here this can happen in the second period due to the arrival of a myth. What happens next depends on how attached players are to the status quo versus how long it takes them to find a new optimum once they are discontent. I am going to give meaning to the status quo by assuming that the length of time it takes to become dissatisfied with the status quo is long relative to how long it takes to find a new optimum once dissatified. That is, content players stick with the status quo until the discontent players settle down. Specifically, the subgroup of players for whom the status quo is no longer optimal eventually become discontent. Once they are discontent they start searching for a new optimum.

The key assumpton is that they reach a *subgroup equilibrium* in which they are all playing best responses to each other and the content players before the content players have a chance also to become discontent.

There are now two possibilities: in this subgroup equilibrium the content players may still be playing best responses, in which case I imagine they remain content and that everyone settles down to that equilibrium, or it may be that some of the previously content players are no longer playing best responses and eventually become discontent. Hence the set of discontent players now consists of those who were previously discontent plus those who found that in the subgroup equilibrium they were no longer playing best responses.

With a new set of content and discontent players, the procedure is then repeated and a new subgroup equilibrium is reached. Eventually a point is reached where either there are some remaining content players left and they are playing best responses in the subgroup equilibrium or all players are discontent. In the former case the content players, who are still playing their status quo actions, together with the subgroup equilibrium constitute a status quo respecting equilibria. In the latter case every Nash equilibrium is status quo respecting.

Roughly speaking this procedure finds a new equilibrium which maximizes the number of players who remain playing their status quo actions. However there can be multiple subgroup equilibria leading to a number of different status quo respecting equilibria depending on which subgroup equilibria occur along the way. In the setting here I will show that this is not the case: there is a unique status quo respecting equilibrium in the second period.

4. Status Quo Respecting Second Period Equilibrium

I begin by characterizing the status quo equilibria. The status quo respecting procedure is simple to implement in the context of the current model and has substantial bite, because it delivers a unique second period equilibrium. Define switching points

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

and

$$\underline{\sigma} = \frac{\Delta - (1 - \phi_L - \phi_H)P}{\gamma}.$$

In the Appendix Theorem 4 shows that

Theorem 1. *For any feasible parameters γ, Δ with a supporting P we have $0 < \underline{\sigma} < \overline{\sigma} < 1$ and there is a unique second period status quo respecting equilibrium of three possible types: 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 $\underline{\sigma}, \overline{\sigma}$.

While the detailed proof is in the Appendix, I want to talk through how status quo respecting equilibrium works. I want to emphasize in particular that for $\sigma \leq \underline{\sigma}$ the status quo remains an equilibrium, so only a modest refinement is required. For other values of σ the status quo respecting refinement guarantees that the middle type and the uninfluenced low and high types, none of whom have any reason to be discontent with the status quo do not abandon it merely because influenced low and high types change their play. Given that these types remain at the status quo, for $\sigma > \overline{\sigma}$ the strong equilibrium is the only equilibrium. In the intermediate range the status quo respecting refinement helps by assuring that the switching of the influenced low types (which must happen) to middle does not cause the high types to suddenly coordinate their behavior and abandon the status quo with which they remain content.

Turning to the specifics, observe that it can only be optimal for influenced types to switch to middle as the arrival of a myth does not change the relative benefit of any other choice. Examining first the influenced high types, their incentive to switch does not depend upon what the low types do. They directly gain Δ by switching to middle. If $\tilde{\phi}_H$ high types remain in the subgroup equilibrium punishment goes from zero to $\tilde{\phi}_H P$ 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, that is, will remain at the status quo, when $\Delta \leq \tilde{\phi}_H P + (1-\sigma)\gamma$. In particular if $\Delta \leq \phi_H P + (1-\sigma)\gamma$ then high types are content and remain at the status quo. On the other hand, if $\Delta > \phi_H P + (1-\sigma)\gamma$ then some influenced high types must switch, so $\tilde{\phi}_H < \phi_H$ and this implies that $\Delta > \tilde{\phi}_H P + (1-\sigma)\gamma$. That is, high types switching make it attractive for other high types to switch, because it reduces the amount of punishment from switching. Hence when $\Delta > \phi_H P + (1-\sigma)\gamma$ all the influenced high types switch to middle. The threshold at which this happens is given by $\overline{\sigma}$.

The incentives of the influenced low types are more complicated because the high types switching make it more attractive to switch to middle as the amount of punishment that can be avoided is greater. Never-the-less, consider first the incentive of the influenced low types when the high types do not switch. A low type loses Δ by switching to middle. Punishment is reduced by $(1-\tilde{\phi}_L-\phi_H)P$ 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 $\Delta \geq (1-\tilde{\phi}_L-\phi_H)P + \sigma\gamma$. As is the case with high types, switching by others of the same type, that is $\tilde{\phi}_L < \phi_L$, only increases the attractiveness of switching. Hence $\underline{\sigma}$ is the cutoff for indifference.

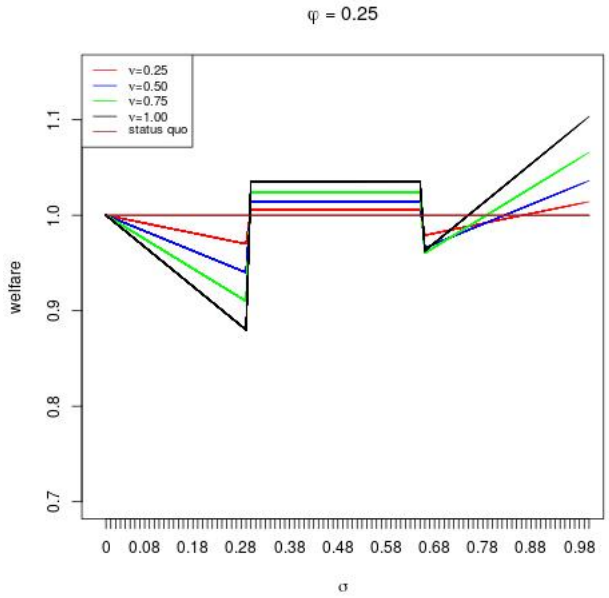
They key remaining fact is that $\underline{\sigma} < \overline{\sigma}$. To see this, observe that this is true exactly when $\gamma + \phi_H P - \Delta > \Delta - (1-\phi_L-\phi_H)P$ which is true if and only if $\gamma > 2\Delta - (1-\phi_L)P$. However, we have assumed $\gamma \geq 2\Delta$, so this is indeed the case. What this means is that when σ is such that the high types

want to switch then the low types want to switch even if the high types do not switch, and certainly want to switch if they do. Hence there are exactly three cases as stated in the Theorem: if $\sigma \leq \underline{\sigma}$ nobody wants to switch and the equilibrium remains at the status quo. If $\underline{\sigma} < \sigma \leq \overline{\sigma}$ then the influenced low types are discontent and switch to middle, but the high types all remain at the status quo. Finally, if $\sigma > \overline{\sigma}$ then both the influenced high and low types are discontent and switch to middle.

5. Myth and Welfare

The main result, Theorem 1 is well illustrated by a class of numerical examples that makes it possible to assess when a myth increases or decreases welfare. I will subsequently examine when the lessons from the example generalizes.

In the numerical examples the status quo is a balanced benchmark with $\phi_L = \phi_R = \phi$. 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 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.

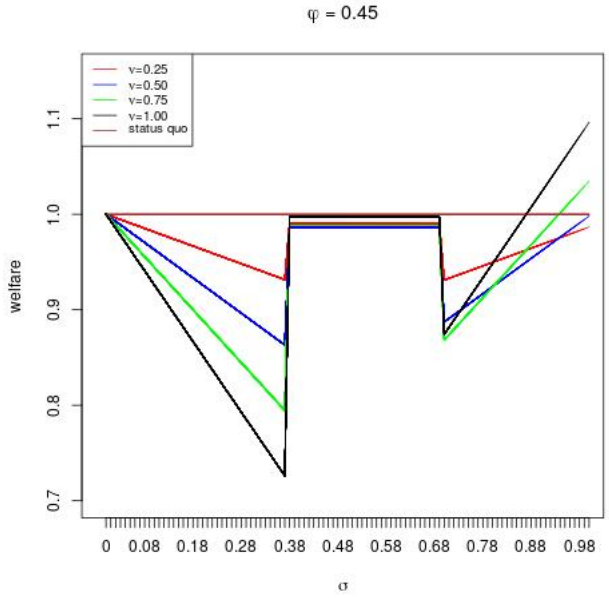
The fact that $0 < \underline{\sigma} < \overline{\sigma} < 1$ means that the influenced low types become indifferent first. As σ increases past the point of indifference, the strict best response is for the influenced low types to switch. Other low types switching increases the incentives of a low type 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 as it is here.

At this moderate equilibrium nobody feels 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. 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%.

6. The Output and Punishment Effects

In the numerical example as persuasiveness σ increases welfare declines linearly, jumps up, is flat, jumps down, then increases linearly. Moreover, when σ passes the second switchpoint, welfare drops below the status quo level. Moderately persuasive and highly persuasive myth are welfare improving only when the status quo is not too polarized. How general are these finding? From Theorem 1 we know that the picture of welfare between the switchpoints is correct: but in general how does welfare jump at the switchpoints? Here I will consider the general case and argue that the picture from the numerical examples is robust.

Theorem 4 in the Appendix computes welfare for all cases. By subtraction we can then find the jumps at the switchpoints $W_m - W_w(\underline{\sigma})$ and $W_s(\overline{\sigma}) - W_m$ as well as the welfare gain over the status quo at the switchpoints $W_m - W_1$ and $W_s(\overline{\sigma}) - W_1$. I have also computed the welfare gain at the top point when there is full persuasiveness ($\sigma = 1$) versus the status quo where there is none. The results of the calculation are reported in the table below. The welfare gain can be decomposed into two parts which I refer to as the output and punishment effects: each is listed on a separate line of the table, and the two are added together to get the total welfare gain.

$W_m - W_w(\underline{\sigma})$	$f(y_1 + \phi_L \Delta\nu) - f(y_1) > 0$
	$-\nu(1-\nu)\phi_L^2 P < 0$
$W_m - W_1$	$f(y_1 + \phi_L \Delta\nu) - f(y_1) - \phi_L \Delta\nu$
	$(1 - \phi_L - \phi_H - (1-\nu)\phi_L)\,\nu\phi_L P$
$W_s(\overline{\sigma}) - W_m$	$f(y_1 - \nu\phi_H\Delta + \phi_L\Delta\nu) - f(y_1 + \phi_L\Delta\nu) < 0$
	$(1 - (1-\nu)(\phi_L + \phi_H))\,\nu\phi_H P > 0$
$W_s(1) - W_1$	$f(y_1 + \nu(\phi_L - \phi_H)\Delta) - f(y_1) - \nu\phi_L\Delta$
	$\left(\phi_L + \phi_H - (2-\nu)\phi_L^2 - (2-\nu)\phi_L\phi_H - (2-\nu)\phi_H^2\right)\nu P$

The first thing to observe is that the jumps do not depend on γ, but then again, neither does welfare at the switchpoints or the top point. A bigger γ means that for $\sigma < \underline{\sigma}$ welfare drops faster but switches sooner, while for $\sigma > \overline{\sigma}$ welfare switches later but rises faster. In brief: the middle range is bigger the larger is γ.

In general the welfare jumps at the switchpoints consists of two offsetting effects. The output effect is driven by the jump in effort $\nu\phi\Delta$ due to $\nu\phi$ extreme types switching to y_M. The punishment effect is due to the the increase in punishment due to low types switching to middle or the decrease due to high types switching to middle.

Notice that at the switchpoints, in contrast to the discrete welfare comparisons with the endpoints, the increased or decreased cost of effort is entirely internalized by the type switching, so only the increase or decrease in output matters. For low types switching output increases so the output effect is positive, while for high types switching output decreases so the output effect is negative. The punishment effect is always the opposite: for low types switching

punishment of goes up and this decreases welfare, while for high types switching punishment goes down and this increases welfare. In other words at the switchpoints while each effect can be signed, they have opposite signs.

The output effect is also complicated by the fact that it is non-linear and we do not know a great deal about f. In the example it was $2\sqrt{x_t}$ but this is obviously special. It is useful to distinguish between the first and second order components of f. I study the first order effect first.

First Order Effects

To first order case we can use $f'(1) = 1$ to approximate $f(x_t) \approx f(1) + (x_t - 1)$. The table below reports the corresponding approximate welfare changes, where I use "$\sim$" to mean "the approximation has the same sign as. Since $\Delta > P$ we can sign some of these as shown. I note also that the final two expression are exact in the balanced case, no approximation is needed.

$W_m - W_w(\underline{\sigma})$	$\sim \Delta - (1-\nu)\phi_L P > 0$
$W_m - W_1$	$\sim 1 - \phi_L - \phi_H - (1-\nu)\phi_L$
$W_s(\overline{\sigma}) - W_m$	$\sim -\Delta + (1 - (1-\nu)(\phi_L + \phi_H))\, P < 0$
$W_s(1) - W_1$	$\sim -\phi_H \Delta + \left(\phi_L + \phi_H - (2-\nu)\phi_L^2 - (2-\nu)\phi_L\phi_H - (2-\nu)\phi_H^2\right) P$

The clear conclusion is that to first order at the switchpoints the output effect always dominates, so the initial jump at $\underline{\sigma}$ is always up and the second jump at $\overline{\sigma}$ is always down. In this sense the numerical example describes a relatively general case.

I turn next to the welfare comparison with the endpoints.

When Are Moderately Persuasive Myths Bad?

Moderately persuasive myths are bad when $W_m - W_1 < 0$. In the numerical example this happened due to polarization. In general according to the first order approximation the output effect does not matter, so moderately persuasive myths are bad exactly when

$$\nu < 1 - \frac{\phi_M}{\phi_L}.$$

This reinforces the conclusion from the numerical example that polarization is bad for moderately persuasive myths in the sense that decreasing ϕ_M without decreasing ϕ_L expands the range of ν for which moderatively persuasive myths are bad. We learn as well that a downward biased status quo, in the sense that ϕ_L is increased at the expense of ϕ_H, has the same effect.

When Are Persuasive Myths Bad?

If $\Delta >> P$ then the output effect dominates and persuasive myths are bad. Since $\Delta > P$ a sufficient condition for $W_s(1) - W_1$ to be negative is

$$\nu < 2 - \frac{\phi_L}{(1-\phi_M) - 2\phi_L\phi_H}.$$

Increased polarization that reduces ϕ_M while not decreasing ϕ_L increases the range of ν for which persuasive myths are bad, again reinforcing the lesson from the numerical example. Morever, holding ϕ_M fixed a downward biased status quo, that is larger ϕ_L and smaller ϕ_H has the same effect.

Second Order Effects

The exact output effect is $f(y_1+\phi_L\Delta\nu)-f(y_1)$ at the lower switchpoint and $f(y_1-\nu\phi_H\Delta+\phi_L\Delta\nu)-f(y_1+\phi_L\Delta\nu)$ at the upper switchpoint. In each case from the fundamental theorem of calculus the discrete difference is the average of slopes between the endpoints. Since f is concave this implies that if the status quo is downward biased, so $\phi_L > \phi_H$, then the output effect is stronger at the first switchpoint than the second and conversely. That is to say, downward bias reinforces the picture in the numerical example about the relative size of the jumps.

In general downard bias strengthens the output effect as f has a steeper slope as effort is reduces, and conversely. In one case this enables us to reach a strong general conclusion. Suppose that the downward bias is strong enough that $(\phi_H-\phi_L)\Delta+\phi_L\Delta\nu<0$, that is, the influenced low types switching is insufficient to reach the first best. In this case $f'>1$ in all cases, reinforcing the output effect. This enables us to conclude that the jump at the lower switchpoint is up and the jump at the lower switchpoint is down as was the case in the numerical example.

7. In the Long Run We Are All The Same

I now want to consider what happens when after the second period additional myths arrive.

I will continue to assume that myths are true myths, with the myth σ_t, ν_t arriving at the end of period t according to Markov process on a finite set $\Sigma, \mathcal{V}$. These sets are non-trival in the sense that there each contains a strictly positive element. There is a common discount factor δ and as is standard in the repeated game literature I will use average present value in order to compare welfare across different discount factors.

I need now need to be more specific about who is influenced: I want to allow for the possibility that some members are more likely to be influenced by myths than others. Specifically, I suppose that in addition to the social norm types there are K equally likely *myth reception* types denoted by $\mu_k > 0$ where $\sum_{k=1}^{K}\mu_k = 1$ and the fraction of reception type k influenced by the myth is $\mu_k\nu_t$. Hence a type with high μ_k is more likely to be influenced by a newly arived myth. Social norm types and myth reception types are independently distributed.

The guilt process remains unchanged. Denote by σ_t^i either σ_t for an influenced member or 0 for an uninfluenced member: guilt evolves according to $g_t^i(x_t^i) = (1-\sigma_{t-1}^i)g_{t-1}^i(x_t^i)+\sigma_{t-1}^i h(x_t^i)$. Note that this means that the effect of repeatedly receiving a myth attentuates: if $g_{t-1}^i(x_t^i)$ is far from the middle type

then guilt adjusts more strongly than if $g^i_{t-1}(x^i_t)$ is close to the middle type. In particular

$$g^i_t(x^i_t) = \left(1 - \left[\sigma^i_{t-1} - \sigma^i_{t-1}\sigma^i_{t-2} + \sigma^i_{t-1}\right]\right) g^i_{t-2}(x^i_t) + \left[\sigma^i_{t-1} - \sigma^i_{t-1}\sigma^i_{t-2} + \sigma^i_{t-1}\right] h(x^i_t)$$

showing both that the order in which the myths arrive is irrelevant and how the arrival of a myth is attenuated by $\sigma^i_{t-1}\sigma^i_{t-2}$.

The relevant notion of equilibrium continues to be open loop. As current behavior has no effect on future utility this means that each member simply optimizes myoptically in each period, so that an open loop equilibrium simply consists of a sequence of Nash equilibria of each period game. The status quo refinement now refers in each period to the status quo from the previous period and the idea of ruling out collective deviations continues to make sense.

For an individual member i we may define $s^i_t = (1-\sigma_t)s^i_{t-1} + \sigma^i_t$ and $g^i_t(x^i_t) = (1-s^i_t)g^i_1(x^i_t) + s^i_t h(x^i_t)$. Depending on the stochastic arrival of myths this gives rise to Q_t, the cdf of s^i_{t-1} at t. I can now describe the equilibrium of the repeated myth game: the proof is in Theorem 5 in the Appendix.

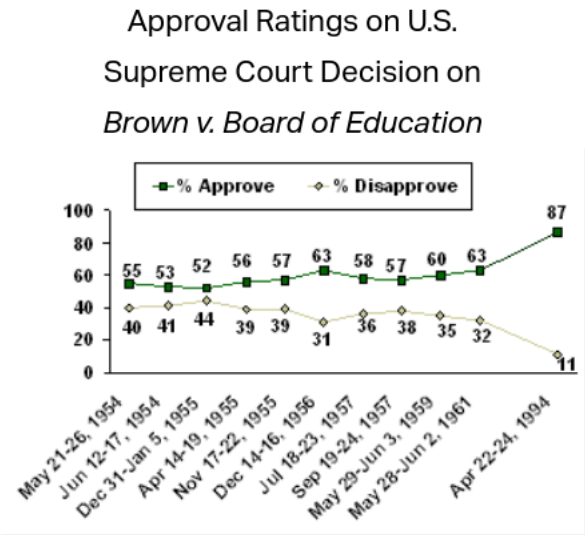
Theorem 2. *There is a unique status quo preserving equilibrium. Let $\hat{\phi}_{Ht}$ be the largest fixed point of $\phi_H = 1 - Q_t(1 - (\Delta - \phi_H P)/\gamma)$ and $\hat{\phi}_L$ the largest fixed point of $\phi_{Lt} = 1 - Q_t((\Delta - (1 - \phi_L - \hat{\phi}_H)P)/\gamma)$. Then there are $\hat{\phi}_{Lt}, \hat{\phi}_{Ht}$ remaining low and high types in period t and this is monotone decreasing. If in addition the myth arrival process is positively recurrent then with probability one $\hat{\phi}_{Lt}, \hat{\phi}_{Ht} = 0$ in finite time and for every $\epsilon > 0$ there is a δ so that the expected present value of welfare converges to within ϵ of the first best.*

This result presents a contrast to the short-run analysis of Theorem 1. That theorem details cases in which true myths may be counterproductive - especially when polarization is high. In the long-run this does not matter, hence in environments, such as those that are highly polarized, in which myth is counterproductive in the short-run it becomes a costly investment towards a good long-run outcome. The extent to which it is worth making this investment depends, naturally, on the discount factor.

Is There Convergence?

Do we reach agreement over social norms in the long run? My impression is that in many cases we do: that issues such as social security, or government measures against air pollution, that were once controversial, are no longer so. Of course, things change, and we find new things to disagree over. As a result the goal posts change and it is not easy to find long series of opinion polls that are relevant to the issue of convergence. One example I did find is a long series of polls concerning the Supreme Court decision in Brown v. Board of Education in which the Supreme Court ruled that discrimination in public education is

unconstitutional. Below I reproduce[2] a time series of polls concerning public approval of that court ruling: I think we may take this as a reasonable proxy for the social norm concerning whether racial discrimination is acceptable or not. As can be seen opinions which initially were about equally split have converged over half a century to near consensus. Notice, incidentally that the adjustment is rather slow: three years after the court decision, opinion has shifted from 55/40 to 57/38. The cost of the conflict was by no means negligible: in 1957, for example, it involved a confrontation between the Arkansas National Guard and the US Army.



It is important to distinguish between disagreement over social norms concerning discrimination, which is what this paper is about, and disagreement concerning the facts of discrimination, which is what the vast literature on information is about. From the polling data I take it that a social norm that was once split between those thinking that discrimination is acceptable and those who feel it is not has converged on an agreement that discrimination is not acceptable. I think there is little doubt that today this social norm is enforced through social sanctions: news stories over people receiving hate mail or losing their jobs, that is, social sanctions, over racial profiling have become commonplace. The story of Amy Cooper[3] who lost her job over a racial incident that occured while she was walking her dog in Central Park is salutory in this respect.

Yet even as the social norms about discrimination have converged, agreement over the facts of discrimination have diverged. In 1954 I do not suppose there was much disagreement over the fact of widespread discrimination. Just to

[2] https://news.gallup.com/poll/11686/race-education-years-after-brown-board-education.aspx

[3] The story is in Wikipedia.

mention a few examples, signs like "White Waiting Room" were common and discrimination in education was explicit policy - as witness the Supreme Court decision in Brown versus the Board of Education. One consequence of the converging social norm that discrimination is bad is not that is has vanished, but that it has become hidden. Hence the question of whether people of color are discriminated against in housing, by the police, and so forth, is no longer obvious and requires a sophisticated analysis of empirical data. As a consequence we now have substantial disagreement about the extent of discrimination. A recent Gallup poll,[4] for example, shows that opinions are about equally split over whether or not black people are discriminated against in jobs and housing.

Cascades

With three or more periods a new phenomenon arises that does not occur in two periods: there can be cascades. In a cascade the switching by types who are influenced by the current myth may trigger switching by types who are not influenced by the current myth. This idea along with the way in which the dynamic equilibrium works can be explained by extending the numerical example to a third period. For simplicity I assume that all members have the same receptivity to influence so that the individual level shocks are, in effect, iid across periods.

Recall that in the numerical example, the parameters were $\gamma = 1.5$, $\Delta = 0.6$, $P = 0.3$, $\phi_{L1} = \phi_{H1} = 0.25$. Here I consider two sequential myths: both have the same pervasiveness $\nu_1 = \nu_2 = 0.5$ but the first is more persuasive $\sigma_1 = 0.64$, $\sigma_2 = 0.27$. To determine the equilibrium it is necessary to compute the utility gain of switching depending on the current value of s_t^i: for remaining low types this is given by

$$(1 - \hat{\phi}_{Lt} - \hat{\phi}_{Ht})P + s_t^i\gamma - \Delta = 0.3(1 - \hat{\phi}_{Lt} - \hat{\phi}_{Ht}) + 1.5s_t^i - 0.6$$

and for remaining high types by

$$\Delta - \hat{\phi}_{Ht}P - (1 - s_t^i)\gamma = 0.6 - 0.3\hat{\phi}_{Ht} - 1.5(1 - s_t^i).$$

The second period parameters are chosen so that the persuasiveness $\sigma_1 = 0.64$ is just less than $\overline{\sigma}_1 = 0.065$. This means that half the low types switch in period 1 while the high types do not, so that the equilibrium is $\hat{\phi}_{L2} = 0.125$ and $\hat{\phi}_{H3} = 0.250$. The third period parameters are chosen so that at the third period status quo the remaining influenced low types slightly prefer not to switch.

However: in the third period, the high types that are influenced in both periods do want to switch, getting a utility gain of 0.1308. This make it more attractive for the other types to switch: the low types because they can avoid more punishment by switching and the high types because they are punished less for switching. In particular the high types that are influenced only in the

[4] https://news.gallup.com/poll/352832/americans-confidence-racial-fairness-waning.aspx

first period switch now get a utility gain of 0.00375 from switching, so do so. However, the high type cascade ends here: even after all the high types who were influenced in the first period switch those who were influenced only in the second do not, getting a utility gain of -0.87. Never-the-less there is now a cascade of low types: after all the high types who were influenced in the first period switch all the influenced remaining low types want to switch, getting a utility gain of 0.18.

In the end all the low types that were influenced in either period switch while all the high types that were influenced in the first period switch, and there remains $\hat{\phi}_{L3} = 0.0625$ and $\hat{\phi}_{H3} = 0.125$.

The theory of cascades presents a contrast to viral network models such as those of Shiller (2017), Shiller (2019) and Benabou, Falk and Tirole (2020). These models lack a notion of social enforcement so have no mechanism in which switching by some in response to a myth alters the punishment landscape and so leads others to switch as well. In contrast to those models, here social norms can spread in leaps and bounds, driven by relatively unimportant myths in an already susceptible population.

8. Conclusion

More or less by assumption pervasive and persuasive myths are always good. In the short-run when polarization is high no other kind of myth is helpful. Otherwise myths that are mediocre on both the pervasive and persuasive front are desirable.

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. By using care of which myths are promulgated the short-run cost of investing in myth can be mitigated.

References

Acemoglu, Daron, and James A. Robinson (2001): "A theory of political transitions," *American Economic Review*: 938-963.

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.

Belloc, M., F. Drago and R. Galbiati (2016): "Earthquakes, religion, and transition to self-government in Italian cities," *The Quarterly Journal of Economics* 131: 1875-1926.

Benabou, Roland, Armin Falk, and Jean Tirole (2020): "Narratives, imperatives, and moral persuasion," University of Bonn.

Block, J. I., D. Fudenberg and D. K. Levine (2019): "Learning Dynamics Based on Social Comparisons," *Theoretical Economics*, 135-172.

Brenneis, Donald (1988): "Telling Troubles: Narrative, Conflict and Experience," *Anthropological Linguistics* 30: 279–91.

Bigoni, Maria, Stefania Bortolotti, Marco Casari, Diego Gambetta, Francesca Pancotto (2013): "Cooperation Hidden Frontiers: The Behavioral Foundations of the Italian North-South Divide," University of Bologna.

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.

Dell, M., Lane, N., and Querubin, P. (2018): "The historical state, local collective action, and economic development in Vietnam," *Econometrica* 86: 2083-2121.

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

Eliaz, Kfir, and Ran Spiegler (2020): "A model of competing narratives," *American Economic Review* 110: 3786-3816.

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.

Foster, D. P. and H. P. Young (2003): "Learning, hypothesis testing, and Nash equilibrium," *Games and Economic Behavior*, 73-96.

Foster, D. P. and H. P. Young (2006): "Regret testing: learning to play Nash equilibrium without knowing you have an opponent," *Theoretical Economics*, 341-367.

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

Fudenberg, D. and D. K. Levine (1992): "Maintaining a Reputation when Strategies are Imperfectly Observed," *Review of Economic Studies* 59: 561-580.

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.

Michalopoulos, Stelios, and Melanie Meng Xue (2021): "Folklore," Quarterly Journal of Economics 136: 1993-2046.

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.

Shiller, Robert J. (2017): "Narrative economics," *American Economic Review* 107: 967-1004.

Shiller, Robert J. (2019): *Narrative economics*, Princeton: Princeton University Press.

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.

Young, H. P. (2009): "Learning by trial and error," *Games and Economic Behavior*, 626-643.

Appendix

Theorem 3. *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. *For any feasible parameters γ, Δ with a supporting P we have $0 < \underline{\sigma} < \overline{\sigma} < 1$.*

In the first period each type contributes their quota. Output is $y_1 = 1 + (\phi_H - \phi_L)\Delta$. Welfare is $W_1 = f(y_1) - y_1 - (\phi_L(1 - \phi_L) + (1 - \phi_L - \phi_H)\phi_H)\, P$.

In the second period there is a unique status quo respecting equilibrium of three possible types: weak, moderate, and strong.

Weak equilibrium occurs for $\sigma \leq \underline{\sigma}$ and every type contributes their first period quota. Output is y_1. Welfare is $W_w = W_1 - \sigma\nu\phi_L\gamma$.

Moderate equilibrium occurs for $\underline{\sigma} < \sigma \leq \overline{\sigma}$. Influenced low types contribute $y_m = 1$ and all other types contribute their first period quota: output is $y_m = y_1 + \phi_L\Delta\nu$, and welfare is $W_m = f(y_m) - y_m - ((1 - \nu)\phi_L(1 - \phi_L + \nu\phi_L) + (1 - \phi_L - \phi_H + \nu\phi_L)\phi_H)\, P$.

Strong equilibrium occurs for $\sigma > \overline{\sigma}$. Influenced types contribute $y_M = 1$ and all other types contribute their first period quota. Output is $y_s = y_1 + \nu(\phi_L - \phi_H)\Delta$ and welfare is $W_s = f(y_s) - y_s - ((1 - \nu)\phi_L(1 - (1 - \nu)\phi_L) + (1 - (1 - \nu)\phi_L - (1 - \nu)\phi_H)(1 - \nu)\phi_H)\, P - (1 - \sigma)\nu\phi_H\gamma$.

Proof. The first period equilibrium is unique and status quo by Theorem 3. There is no guilt, output is the first best by construction, so welfare is $f(y_1) - y_1$ minus the cost of punishment. The low type is punished by the middle and high

type, so receives an punishment of $(1-\phi_L)P$. The middle type is punished by the high type so receives an expected punishment of $\phi_H P$. The high type is not punished. As there are ϕ_L low types and $1-\phi_L-\phi_H$ middle types the average expected punishment is

$$\phi_L(1-\phi_L)P + (1-\phi_L-\phi_H)\phi_H P$$

giving welfare as indicated.

The types of status quo equilibria in the second period and the switchpoints were established in the text along with the fact that $\overline{\sigma} > \underline{\sigma}$. I show that the switchpoints in fact satisfy the property of lying strictly between 0 and 1. For the upper switchpoint $\overline{\sigma} > 0$ follows from $\gamma \geq 1$ and $\Delta < 1$. Since Theorem 3 requires that $\Delta \geq P > \phi_H P$ we also have $\overline{\sigma} < 1$. For the lower switchpoint Theorem 3 requires that $\Delta \geq P > (1-\phi_L-\phi_H)P$ so $\underline{\sigma} > 0$, and from $\Delta < 1$ and $\gamma \geq 1$ it follows that $\underline{\sigma} < 1$.

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 $\nu\phi_L\Delta$ as asserted. Nobody feels guilty in this equilibrium, so it remains to work out the punishment cost. After switching there are $(1-\nu)\phi_L$ low types and $(1-\phi_L-\phi_H+\nu\phi_L)$ middle types. The former are punished all other types, so get an expected punishment of $(1-\phi_L+\nu\phi_L)P$. The latter are punished by the high types so get an expected punishment of $\phi_H P$. Averaging with the number of types of each gives welfare as indicated.

In the strong case, punishment cost is determined by replacing ϕ_τ with $(1-\nu)\phi_\tau$ for $\tau = L, H$ 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. □

Theorem 5. *There is a unique status quo preserving equilibrium. Let $\hat{\phi}_{Ht}$ be the largest fixed point of $\phi_H = 1 - Q_t(1-(\Delta-\phi_H P)/\gamma)$ and $\hat{\phi}_L$ the largest fixed point of $\phi_{Lt} = 1 - Q_t((\Delta-(1-\phi_L-\hat{\phi}_H)P)/\gamma)$. Then there are $\hat{\phi}_{Lt}, \hat{\phi}_{Ht}$ remaining low and high types in period t and this is monotone decreasing. If in addition the myth arrival process is positively recurrent then with probability one $\hat{\phi}_{Lt}, \hat{\phi}_{Ht} = 0$ in finite time and for every $\epsilon > 0$ there is a δ so that the expected present value of welfare converges to within ϵ of the first best.*

Proof. The middle types always choose the middle quota so the interest is only in the low and high types. For the high tpes the indifference point is $\Delta = \phi_H P + (1-s)\gamma$ where $\phi_H = 1 - Q_t(s)$ giving the Nash equilibrium values of ϕ_H. Status quo respecting demands the largest fixed point. Once $\hat{\phi}_{Ht}$ is known, for the low types the indifference point is $\Delta = (1-\phi_L-\hat{\phi}_H)P + s\gamma$ where $\phi_L = 1 - Q_t^L(s)$ giving the Nash equilibrium values of ϕ_L, and again status quo respecting demands the largest fixed point.

For the asymptotic part of the result, define $r_t^i = 1 - s_t^i$ and $\underline{r}_t$ as the minimum of r_t^i at time t. Since $r_t^i = r_{t-1}^i - \sigma_t^i r_{t-1}^i$ it is non-increasing, so $\underline{r}_t$ is as well. If

$\underline{r}_t < 1-\overline{\sigma}$ then everyone has switched and will remain switched forever. Let $\overline{\Sigma}, \overline{\mathcal{V}}$ be the maximal elements of $\Sigma, \mathcal{V}$: by assumption they are positive. Let $\underline{\mu}$ be the smallest value of μ_k also positive by assumption.If the myth arrival process is positively recurrent then there is a T and a $\lambda > 0$ such that over T periods the myth $\overline{\Sigma}, \overline{\mathcal{V}}$ arrives. Since $r_t^i/r_{t-1}^i - 1 = -\sigma_t^i$ when this occurs $|r_t/r_{t-1} - 1| \geq \overline{\Sigma}$ with probability at least $\overline{\mathcal{V}}$. If we measure r_t at times $T = T, 2T, 3T, \ldots$ the process adapted to the filtration induced by the underlying myth arrival process is a supermartingale with the additional process that it is active, meaning that the probability it jumps by a minimum amount is bounded away from zero. These supermartingales are studied in the Appendix to Fudenberg and Levine (1992) who showed that not only do they converge to zero with probability one, establishing that $\hat{\phi}_{Lt}, \hat{\phi}_{Ht} = 0$ in finite time, but they do so at a uniform rate in the sense that for any $\epsilon, r > 0$ there is a K such that $\Pr(\sup_{t>KT} r_t \leq r) < 1-\epsilon$. Choosing $r < 1 - \overline{\sigma}, \epsilon(1 - \delta^{KT})$ and observing that the worse utility a member receives in a period is $\underline{u} = -1 - \Delta - \gamma - P$ this means that the expected average present value received by any member is $(1-\delta^{KT})(1+\epsilon)\underline{u}+\delta^{KT}(1-\epsilon)[f(1)-1]$. As $\delta \to 1$ this converges to $(1-\epsilon)(f(1)-1)$ as asserted.

so that $g_t^i(x_t^i) = (1 - s_{t-1}^i)g_1^i(x_t^i) + s_{t-1}^i h(x_t^i)$ where s_{t-1}^i depends upon the myths influenced by but not the order in which they are received □

high enough discount factor first best present value: relevance to patience of sender?