Moral Narratives and Social Norms

Roland Bénabou[1], Armin Falk[2], Jean Tirole[3]

April 6, 2026[4]

[1]Princeton University, NBER, CEPR, CIFAR, IZA BREAD, THRED and ERINN.

[2]University of Bonn.

[3]Toulouse School of Economics (TSE) and Institute for Advanced Study in Toulouse (IAST), University of Toulouse Capitole.

[4]An earlier and longer version of this paper was titled "Narratives, Imperatives, and Moral Persuasion." We are thankful for valuable comments and suggestions from Johannes Abeler, Alberto Alesina, Nageeb Ali, Daniel Chen, Alexander Dorofeev, Thomas Graeber, Johannes Hermle, Ian Jewitt, Alessandro Lizzeri, Pietro Ortoleva, Stephen Morris, Christopher Roth, Gilles Saint-Paul, Paul Seabright, Nora Szech, Adam Szeidl, Joël van der Weele, Leeat Yariv and participants at many seminars and conferences. Ana Luisa Dutra, Juliette Fournier, Pierre-Luc Vautrey, Thorben Woelk and Ben S. Young provided superb research assistance. Bénabou gratefully acknowledges financial support from the Canadian Institute for Advanced Study, Tirole and Falk from the European Research Council (European Community's Seventh Framework Programme Grant Agreement no. 249429 and no. 340950, as well as European Union's Horizon 2020 research and innovation programme, Grant Agreement no. 669217).

Abstract

We study the production and circulation of arguments justifying actions on the basis of morality. By downplaying externalities, exculpatory narratives allow people to maintain a positive image while acting selfishly. Conversely, responsibility narratives raise both direct and reputational stakes, fostering prosocial behavior. These rationales diffuse along a linear network, through both costly actions and strategic disclosure. The norms that emerge reflect local correlation in agents' incentives (reputation versus influence concerns), with low mixing generating both a polarization of beliefs across groups and less moral behavior on average. We also study agents' own search for narratives and show how this can lead to lenient or strict moral standards for how strong excuses must be in order to be admissible.

Keywords: Moral behavior, narratives, imperatives, rules, excuses, responsibility, networks, viral transmission, influence, reputation, disclosure, communication, social norms.
JEL Codes: D62, D64, D78, D83, D85, D91, H41, K42, L14, Z13

1 Introduction

1.1 Moral decisions and narratives

Appeals to moral responsibility are central to public goods provision and the upholding of norms, while rationales for acting according to self-interest undermine them. We refer to such arguments as *moral narratives* and study how they propagate through society and shape behaviors. We also examine agents' incentives to come up with such rationales on their own, and how strong excuses must then be in order to be admissible.

We use the term "narrative" because people's beliefs about the morality of different behaviors can be affected not just by objective facts but also by suggestive stories or cues, ex-post rationalizations and interpretations, stereotypes, emotional manipulations and other "behavioral" modes of persuasion that play on cognitive limitations or motivated reasoning. For our purposes, all that matters is that these messages be persuasive and can then be passed on to someone else.[1]

Two broad types of narratives are relevant for moral behavior. By downplaying externalities or emphasizing personal costs, *exculpatory narratives* allow an individual to behave selfishly while maintaining a positive self- and/or social image. Common examples include denials of harm ("it wasn't that bad"), of responsibility ("I had no choice,", "I was not pivotal, we just followed orders") or moral significance ("everyone does it"), and the derogation of victims. Conversely, *responsibility narratives* increase the pressure to behave prosocially, which is costly. Classical moralizing arguments involve appeals to empathy ("how would you feel in their place?"), universalizing counterfactuals ("what if everyone did this?", "think about the planet"), and higher moral authorities (religion, role models).

Incentives, network structure, and virality. The first question we take up is: what types of individual motives and social interactions lead exculpatory versus responsibility rationales to spread more widely, or remain clustered within subgroups?

The main ingredients of the model are as follows. Following the utilitarian tradition, we define an action as moral if it produces a positive externality, or averts a negative one. A population of agents chooses between taking such a prosocial action, at some personal cost, or a selfish one that is costless. Their concern for others' welfare can be high or low, and they derive reputational benefits from being perceived (or seeing themselves) as highly moral types. The magnitude of the externality is a priori unknown but agents may learn a narrative about it, either exogenously (say, through the media) or from someone else who relays it to them. We formalize this communication as (equivalent to) the disclosure of a hard signal about the externality, drawn from some given distribution. It can be taken literally as such; but, as mentioned above, all that matters is that the narrative affect the beliefs of any agent receiving it. Agents interact along a simple linear (or tree) network that stochastically mixes individuals with different signaling and disclosure incentives, whom we call "active" and "passive". Each agent may observe what their predecessor did, receive a narrative either from them or exogenously, and then transmit it, or not, to their successor. Active agents decide whether to take the moral action or not and what narratives to share or withhold, both of which will affect their reputation in the eyes of their successor, and potentially his behavior. Passive agents can also

[1] Narratives are "instruments of mind in the construction of reality" (Bruner 1991).

exert influence through what arguments they convey (or not) to their successor, but their own morality is not at stake: either they have no externality-generating material action to take, or they always want to behave prosocially. Our running example is that of a dominant or majority ethnic group and a less powerful or minority group, how the former treat the latter, and the different arguments circulating about those behaviors. Other applications include men and women, or rich and poor in the context of transfers. We obtain three sets of results.

First, the spread of different rationales through a population is driven by type-specific tradeoffs between *reputation* and *influence* motives. For instance, an actor who learns of a narrative justifying selfish behavior has a social-image incentive to share it with his observer-successor. If he does so, however, the latter now has the excuse on hand, making him more likely to act similarly and justify it to his own audience, and so on –a negative multiplier. Conversely, sharing a responsibility narrative creates for the sender a pressure to act morally or else face strong stigma, but it has the now positive multiplier effect that the successor may not just act well but also pass on the duty argument to his next neighbor, etc. We show that negative disclosures are strategic substitutes while positive ones are strategic complements, and characterize what individuals relay or withhold depending on their moral types and reputational stakes.

Second, we determine how far each type of narrative travels as a function of the interaction structure, and how this affects overall behavior. The social norm that emerges as a result can be one in which either prosociality or self interest is the default, namely what a moral person is expected to do unless they produce a good argument to the contrary. In the first kind of equilibrium, doing the right thing "goes without saying," whereas abstaining requires an excuse to avoid reputational damage; thus, only negative narratives are used, when available. In the second kind, inaction is the default, so someone pursuing self-interest can plead ignorance, though having an excuse is better, so those again circulate. But now so do positive narratives, propagated by high-morality actors as well as non-actors, both seeking to induce agents down the line to behave responsibly; conversely, for such narratives, intentional "silence is complicity".

Third, we show that in either type of equilibrium, more *mixed interactions* between agents with differing reputation-influence tradeoffs *raise prosocial behavior.* Passive agents, whose morality is not at stake, have no need for excuses, so they act both as "firewalls" limiting the spread of exonerating narratives and as "relays" for responsibility ones. In the latter case so do high-morality actors, with complementary amplification. As a result, moral behavior is maximized when active and passive types alternate along the line, a negative serial correlation. Turning from average behavior to dispersion, we show that the degree of clustering and *polarization* of beliefs is minimized by random mixing (zero correlation). Conversely, a significant correlation causes opposite narratives to circulate in the two groups. In the lead example, the majority group will share *more excuses* and rationalizations for behaviors that the minority will simultaneously view as *more inexcusable,* compared to what would occur in a more integrated setting.

Moral standards. Having emphasized the social transmission of arguments, we turn to the "production" side, by allowing an individual to engage in his own *search for reasons to act,* or not. Formally, the probability of drawing a signal or narrative about the externality is an increasing function of search intensity, which entails information or cognitive costs. A new question then arises, namely how the mere fact that someone has a justification for self-

interested behavior should be interpreted. Is it more indicative of a low-morality type who only looks for excuses, or a high-morality one who seeks to ascertain his responsibilities and found out that the externality is low? The answer depends, intuitively, on the relative search intensities of the two types, which in turn hinges on a comparison between the option value of discovering that the externality is substantially away from the prior mean (in either direction), versus that of learning that it is simply low enough to provide an acceptable excuse. We show how, in equilibrium *moral standards* emerge that specify *how strong excuses must be* to be deemed acceptable, and how much stigma someone offering a "lame" one would incur. We characterize when the equilibrium standard is strict or lenient as a function of the mean and tail moments of the prior distribution of narratives, and show that over some range the strict and stringent standards equilibria can coexist, associated respectively with norms of acting prosocially or according to self interest when uninformed.

1.2 Literature review

The paper ties into several lines of work. The first is the literature linking prosocial behavior to signaling or moral-identity concerns (e.g., Bénabou and Tirole 2006, 2011a,b, Ellingsen and Johannesson 2008, DellaVigna et al. 2012, Exley 2016, Bursztyn et al. 2019). To the usual choice dimension of agents taking some costly prosocial action, we add the sharing of arguments and justifications. Indeed, it is both what people do and what they say that determines how others judge them and respond, and thus how social norms are upheld.

Most closely related is parallel and independent work by Foerster and van der Weele (2021). In their model, a sender with private information about the impact of a prosocial action sends a cheap-talk message about it to a receiver. Both then act, with the receiver observing the action of the sender, who cares about being perceived as prosocial. This image concern creates an incentive to understate the externality, while the desire to spur contributions pushes towards exaggerating it, as in our model where the reputation and influence motives lead to selective disclosure of different narratives. Experimentally, Foerster and van der Weele find evidence of both effects –in particular, raising the visibility of the sender's actions makes them more likely to report low impact, reducing receiver donations. The two models have different communication technologies, and we analyze interactions between many agents: viral diffusion, group polarization, and the effects of social mixing.

The communication aspect of the paper also relates it to strategic information transmission in networks (e.g., Hagenbach and Koessler 2010, Galeotti et al. 2013, Ambrus et al. 2013, Bloch et al. 2018), but our agents communicate through two channels: costly signaling of their private types and selective disclosure of a common state. The role of being pivotal also links it to work on moral responsibility in groups or markets (e.g., Bartling and Fischbacher (2012), Falk and Szech 2013, Falk et al. 2020, Dewatripont and Tirole (2024)). Campbell et al. (2025) model behavioral diffusion on a random network where agents are linked not through information but through an adoption externality. As here, they study when public-good or public-bad actions will be strategic complements or substitutes and how far each will spread. This is shown to depend critically on the network's density, through agents' expected influence (how many other individuals their choice will be solely pivotal for), which generalizes the influence multiplier (expected depth of a cascade one can trigger) in our simpler, linear framework.

Finally, the paper belongs to the fast-growing literature on the role of narratives in economics. The importance of "stories" in shaping people's beliefs was first emphasized by Akerlof and Shiller (2015) and Shiller (2017).[2]

On the theoretical side, a line of work initiated by Spiegler (2016) represents narratives as directed acyclic graphs encoding agents' causal beliefs about the world. Agents adopt the one that best fits observed long-run correlations, generating beliefs that are internally consistent yet potentially biased. Charles and Kendall (2022) provide experimental support for the model, while Eliaz and Spiegler (2020) and Besley and Brzezinski (2025) extend the analysis to competing narratives, and Eliaz and Spiegler (2024) to the strategic supply of narratives by news platforms. A second strand, initiated by Schwartzstein and Sunderam (2021), formalizes narratives as likelihood functions: a receiver exposed to competing models adopts the one that fits the data best given their prior. This gives senders an ability to strategically persuade, which is limited by the fact that a better-fitting model necessarily moves beliefs less away from the prior. Barron and Fries (2023) provide experimental support for the theory, as do Aina and Schneider (2025). Aina (2023) extends it to a persuader who tailors their narratives to influence agents with private signals, and Schwartzstein and Sunderam (2024) to agents in a social network who share their available narratives.

An important channel through which narratives operate is memorability, driven by cues and associations, as demonstrated by Andre et al. (2022) for beliefs about macroeconomic mechanisms and by Graeber, Roth, and Zimmermann (2024) for beliefs about goods and venues. Another channel is media exposure, as documented by Andre et al. (2026). As to where narratives originate, Michalopoulos and Xue (2021) link preindustrial societies' folklore themes to contemporary attitudes and beliefs, while Mukand and Rodrik (2018) share this paper's feature that some agents engage in costly search for narratives, which they then use strategically.

Concerning moral narratives specifically, Dal Bo and Dal Bo (2014) find that both utilitarian and deontological messages significantly raise contributions in a public-good game. In a field experiment Barron et al. (2023) show that narratives about shared family history substantially reduce intergroup discrimination between Jordanian host and Syrian refugee children. Bartling et al. (2024) show that public discourse can shift market outcomes toward social responsibility by changing shared expectations about what kinds of transactions others view as appropriate. Hillenbrand and Verrina (2022) test and generally confirm some of the main implications of our model. Positive narratives increase charity giving, especially by selfish types, and worsen the image of those who do not give, as they make observers judge giving as more socially appropriate. Negative narratives conversely make it seem as less appropriate and shield the image of non-givers. However, they do not lead to less aggregate giving, a result which the authors conjecture might reflect subjects' not wanting to look "influenceable" by justifications for acting selfishly.

Besides narratives and other arguments that convey information about the consequences of an agent's actions (or operate as if they did), another form of moral discourse is *imperatives*, which are entirely soft messages of the type "thou shalt (not) do this," seeking to constrain behavior without offering any reason other than a tautological "because I say so". In Bénabou, Falk, and Tirole (2019) we analyze the costs and benefits of imperatives relative to narratives

[2]For critical surveys of the economics literature on narratives from interdiscipinary perspectives, see Juille and Jullien (2017) and Roos and Reccius (2024).

in shaping moral behavior, and when the two will be used as substitutes or complements.

The paper proceeds as follows. Section 2 introduces a basic setup in which moral values, esteem concerns and beliefs about the externality jointly shape individual behavior. Section 3 embeds it into a linear stochastic network to study how the diffusion of arguments, the resulting norm, and belief polarization reflect the interplay of reputation and influence motives, together with the degree of social mixing. Section 4 turns to the search for narratives and how this leads to equilibria with different standards for what excuses are considered acceptable, and associated behavioral norms. The main proofs are in the main Appendix, auxiliary ones in the Supplementary Appendix.

2 Basic Model

2.1 Moral decisions and types

1. Preferences. There are three periods, $t = 0, 1, 2$. At date 1, a risk-neutral individual will choose whether to engage in moral behavior ($a = 1$) or not ($a = 0$). Choosing $a = 1$ is prosocial in that it involves a personal cost $c > 0$ but may yield benefits for the rest of society, generating an expected externality or public good $e \in [0, 1]$; for instance, e may be the probability of an externality of fixed size 1.

Agents differ by their intrinsic prosocial orientations: given e, it is either $v_H e$ (high, moral type) or $v_L e$ (low, immoral type), with probabilities ρ and $1 - \rho$ and $v_H > v_L \geq 0$; the average type will be denoted as $\bar{v} = \rho v_H + (1 - \rho) v_L$.

In addition to intrinsic satisfaction, acting morally confers a social or self-image benefit, reaped at date 2. In the social context, the individual knows his true type but the intended audience (peers, employers, potential mates) does not. Alternatively, the concern may be one of self-signaling: the agent has a "visceral" sense of his true values at the moment he acts, but later on the intensity of that emotion or insight is no longer perfectly accessible; only the decision itself can be reliably recalled. Either way, an agent of type $v = v_H, v_L$ seeks to maximize

$$U = (ve - c)\, a + \mu \hat{v}(a), \tag{1}$$

where $\hat{v}(a)$ is the expected type conditional on $a \in \{0, 1\}$ and $\mu \geq 0$ measures the strength of self- or social-image concerns, common to all agents.[3] His utility level could also include direct benefits (if any) received from others' decisions, but he takes those as given.

2. Behavior under common knowledge. We first consider the basic situation where actor and audience share the same expectation e about the externality. To limit the number of cases, we make an assumption ensuring that the high type always contributes when the externality is large enough or sufficiently certain, while the low type never does.

[3] It may seem that we only consider here behaviors that both actor and audience judge prosocial, and which the latter accordingly rewards with esteem $\mu \geq 0$: helping, not stealing from or exploiting others, etc. Actions that are judged as moral by one group and immoral by another (abortion, guns, religion, politics, in-group favoritism, etc.) generate strong incentives for assortative matching; if this results in agents with antithetical values having little social contact, we are back to (1). When sorting is imperfect, signaling will involve multiple audiences (see Tirole 2026), but we show in the Appendix how such cases still reduce to an "on net" unidimensional model.

Assumption 1.

$$v_L - c + \mu\,(v_H - v_L) < 0 < v_H - c + \mu\,(v_H - \bar{v})\,. \tag{2}$$

The first inequality says that $a = 0$ is a strictly dominant strategy for the low type: he prefers to abstain even when the social and reputational benefits are both maximal, $e = 1$ and $\hat{v}(1) - \hat{v}(0) = v_H - v_L$. The second inequality says that both types pooling at $a = 0$ is not an equilibrium when the externality is maximal ($e = 1$): the high type would deviate to $a = 1$, even at minimal image gain $v_H - \bar{v}$. When $a_H = a_L = 0$ is an equilibrium, we set $\hat{v}(1) = v_H$, by elimination of strictly dominated strategies. These assumptions also imply that, when the externality is in some intermediate range, multiple equilibria coexist: if

$$v_H e - c + \mu(v_H - \bar{v}) \leq 0 \leq v_H e - c + \mu(v_H - v_L),$$

there exist both a pooling equilibrium at $a = 0$ and a separating equilibrium in which the high type contributes, with a mixed-strategy one in-between. Intuitively, if the high type is expected to abstain there is less stigma from doing so, which in turn reduces his incentive to contribute. In case of multiplicity, we select the equilibrium that is best for both types, namely the no-contribution pooling equilibrium. Indeed, the separating equilibrium yields lower payoffs: $\mu v_L < \mu\bar{v}$ for the low type and $v_H e - c + \mu v_H \leq \mu\bar{v}$ for the high one.[4] Since $v_L \geq 0$, Assumption 1 and our selection criterion imply that the moral type contributes if and only if $e > e^*$, where e^* is uniquely defined by

$$v_H e^* - c + \mu(v_H - \bar{v}) \equiv 0. \tag{3}$$

Intuitively, selfish behavior is encouraged by a low perceived social benefit e, a high personal cost c, and a weak reputational concern μ.[5]

2.2 Moral narratives: exoneration and responsibility

When actor and observer do not share the same belief about the externality e, arguments about its importance will come into play.

Definition. A (moral) *narrative* or argument is any signal or message –whether hard information, frame, cue, rhetorical device, etc.– that, when received by an agent, will move his expectation of the externality from the prior mean e_0 to some other value e, distributed ex-ante on $[0, 1]$ according to a cdf $F(e)$. At any given time there is a single realized narrative e (e.g., piece of real or fake news) that may circulate among agents, but we will study its effects for all values, as well as on average across them (e.g., across periods).

An individual who has learned e (or a signal inducing that posterior mean) can disclose it or not to his audience, $d \in \{e, \emptyset\}$, thereby potentially affecting both how his action $a \in \{0, 1\}$ will be judged and the behavior of others who will learn e from his message, either directly or

[4] Pareto dominance is understood here as better for both types of a single individual. Depending on whether the externalities from $a = 1$ fall on the same set of agents whose actions are being studied or on some outside ones (the poor, countries most vulnerable to global warming, distant generations, other species, etc.), this may be different from (even opposite to) that of making everyone in society better off. If we instead selected the separating equilibrium the main comparative statics of interest would remain the same, however.

[5] When $e > e^*$, the separating equilibrium $(a_H, a_L) = (1, 0)$ is the unique one. When $e \leq e^*$, the pooling equilibrium $(a_H, a_L) = (0, 0)$ exists and is best for both types, and thus selected by our Pareto criterion.

via multiple intermediaries. Denoting by N_+ and N_- the number of decisions thus influenced in the prosocial and selfish directions respectively, utility becomes

$$U = v\hat{e}\left[a + N_+(a,d) - N_-(a,d)\right] - ca + \mu\hat{v}(a,d), \tag{4}$$

where $\hat{e} \equiv e$ and $d \in \{e, \emptyset\}$ for someone who knows e and $\hat{e} = e_0$ and $d \equiv \emptyset$ for someone who is uninformed.[6] We next provide examples of two main types of narratives, then in their light discuss the formal definition.

(1) *Absolving (negative) narratives* serve to legitimize selfish or even intentionally harmful actions, by providing excuses and rationalizations of such acts as consistent with the standards of a moral person. They operate through various exculpatory or neutralization strategies (e.g., Sykes and Matza 1957) such as: (a) downplaying the harm**;** (b) blaming or dehumanizing the victims; (c) denying agency or responsibility (e.g., "I was just following orders") or underestimating being pivotal, as in the bystander effect (Darley and Latane 1968); (d) appealing to higher loyalties like religious values or state interests that justify hurting others in the name of "a greater good."

(2) *Responsibility (positive) narratives,* on the contrary, create pressure to behave well, by emphasizing how a person's actions impact others, as well as the moral duties and inferences that result from such agency: making a difference, setting a precedent, etc. include: (a) Kantian-like universalizations ("What if everyone did the same?" "Do unto others..."); (b) arguments inducing empathy ("What if it were you?") and making salient the plight of others (identifiable-victim effect); (c) moral and religious parables, inspiring myths or role models; (d) stressing common identities, such as national and religious brotherhood, or sharing the same planet; (e) invoking some higher moral authority that will pass judgment (God, Adam Smith's "impartial spectator").

Discussion. We next comment on our strategy of modeling narratives as signals inducing posteriors drawn from $F(e)$. First, low realizations of e will clearly be "negative narratives" or "excuses," while high ones will be "positive narratives" or "responsibilities". How high or low they have to be (compared to e^*) to alter inferences and behavior (the functions $\hat{v}$, N_-, N_+) will be determined in equilibrium.

Second, as many of the examples show, stories need not be objectively true to influence people's behavior and judgment (Haidt et al. 2009). They could be any of: (a) genuine, hard facts accompanied by a correct interpretation; (b) true but selective facts from which people will draw incorrect conclusions, due to systematic biases: framing and salience effects, confusing correlation with causation, base-rate neglect, similarity-based reasoning, etc.; (c) baseless or illogical arguments that strike an emotional chord, or play into wishful thinking. There is now a large literature analyzing various "behavioral" channels through which updating may be distorted, signals ignored over/underweighted or misinterpreted, etc. We purposedly abstract here from choosing any particular channel through which narratives may persuade, to focus instead on *why and how people use them,* as a social equilibrium. The essential feature for any *positive* analysis is that these stories or messages "work" –be subjectively perceived by

[6] We add an infinitesimal disclosure cost to break cases of indifference. If agents do not fully internalize their influence on followers' actions, one just scale downs the terms N_+ and N_-.

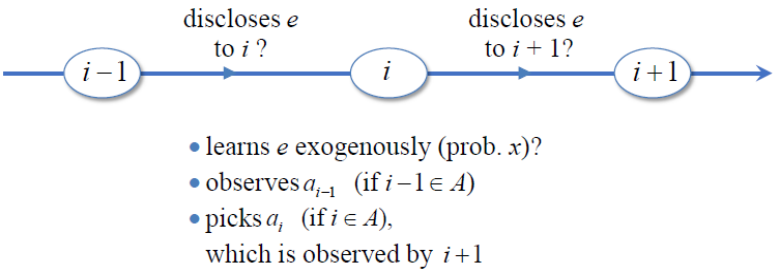

Figure 1: Viral Transmission of Narratives

recipients as containing enough of a "grain of truth" to affect their inferences and behaviors.[7] For *normative* conclusions their veracity does matter, but only in the sense that any equilibrium outcome they generate should be evaluated according to whatever value of e the social planner deems to be the right one.

Third, good arguments are, by definition, scarce: they must be intuitive, salient, memorable, preferably novel and yet consistent with recipients' priors, past experiences, and motivated beliefs. The ex-ante distribution $F(e)$ captures the relative *availability* or/and *persuasiveness* of more or less prosocial ones in a given economic, informational and psychological environment.

We will explore two channels through which narratives and social norms shape each other. Our primary emphasis is on the *social-transmission* channel, studied in the next section. Subsequently, we examine the *production* channel, namely agents' search for arguments.

3 Viral Narratives

"Reasons and arguments can circulate and affect people, even if individuals rarely engage in private moral reasoning for themselves." (Haidt 2001, p. 828-829)

We analyze here the different mechanisms through which exculpatory versus responsibility arguments can spread through a population, and how far each will ultimately travel.

Consider first a negative rationale. An agent who learns of it has an incentive to disclose this excuse to observers, so as to dampen their unfavorable inferences concerning his morality if he chooses to behave selfishly. This *reputational* motive is potentially counterbalanced by a second, *social influence* one: when audience members are themselves actors confronting similar choices, sharing one's excuse with them tends to corrupt their behavior, thereby amplifying the negative externality on society. Even when he is not materially affected by the latter, agent i cares intrinsically about the harm caused by his words (disclosure), just like he cares about that caused by his deeds (action): though one is direct and the other indirect, he is responsible for both. The same reputation and influence effects operate in reverse for positive narratives: sharing information suggesting that some action imposes significant social harm places one's reputation at stake, but the social-influence effect is now positive, as awareness of consequences

[7] Some of the most successful narratives are even demonstrably wrong: conspiracy theories, pseudo-scientific denials of global warming, and other "alternative facts." Barrera et al. (2020) find that incorrect facts embodied in a compelling narrative have a much stronger influence on voting intentions than actual ones, and that correcting the facts does nothing to undo these effects.

promotes others' moral behavior.[8]

3.1 Signaling and disclosure on a linear network

1. Setup. There is a countable set of individuals $i \in \mathbb{Z}$, arranged on a line, who differ in their reputational and influence concerns. Specifically, each can be of one of two activity types: "Passive" , in which case he has no opportunity to act along the dimension of interest, and this is known to his successor $i+1$; or "Active," meaning that he chooses some $a \in \{0,1\}$ and that this action is observed by $i+1$. Equivalently, each active or passive agent could have $k > 1$ successors, but only one predecessor, so that the network is a tree.

Whether active or passive, if someone knows of a narrative e he has a choice of communicating it, or not, to his successor, $i+1$; see Figure 1. An agent does not know whether his successor is active or passive –i.e., exactly who will learn of his words and deeds, but only that types are determined according to a symmetric Markov transition process with persistence $\lambda \in [0,1]$:

$$\Pr[i+1 \in A \mid i \in A] = \Pr[i+1 \in P \mid i \in P] = \lambda, \tag{5}$$

where A and P respectively denote the sets of active and passive individuals.[9] In equilibrium, agents in those two sets will typically have different disclosure strategies, so that what i knows about the externality e will depend on whether $i-1$ was active or passive. The following "time symmetry" implication of (5), resulting from the fact that the invariant distribution of types is 50-50, will therefore be useful:[10]

$$\Pr[i-1 \in A \mid i \in A] = \Pr[i-1 \in P \mid i \in P] = \lambda. \tag{6}$$

Agents' moral preferences remain unchanged: a proportion ρ has type v_H and the remaining $1-\rho$ type v_L, with $v_H > v_L \geq 0$, and all (active) agents share the same reputational concern μ with respect to their audience, which for i is simply their successor $i+1$. As explained earlier, each individual's preference v applies to any externality he causes, whether through a_i (for A types), or by sharing or withholding a narrative (for both A and P types). Preferences are thus given by (4), with $a \equiv 0$ (or $a \equiv 1$) for passive agents.[11]

The distribution of potential signals or narratives will, for simplicity, be taken in this section to be binary: e equals e_- (probability f_-) or e_+ (probability f_+), with $f_+ e_+ + f_- e_- = e_0$ and

[8] That sharing a negative signal (low e) is beneficial to one's reputation, and sharing a positive one (high e) detrimental to it, is a general insight, not limited to the case of selfish choices, $a = 0$. Since intrinsic motivation is ve, choosing $a = 1$ is a stronger signal about v, the lower is e. With only two types and preferences satisfying (2) such inferences do not come into play, as $a = 1$ fully reveals the high type, but more generally they would.

[9] "The" successor of i is thus, in practice, the *set* of individuals who will see what he did and/or hear what he says (including via email, social media, etc.), and λ is the expected fraction with A/P type similar to his. As mentioned above, this stochastic tree structure is isomorphic, for our purposes, to a line with random successors.

[10] Groups A and P differing in size could be accommodated by making the Markov chain asymmetric. Another (more complex) extension would involve P's taking another action that symmetrically affects the A's, or/and also affects other P's. It should be very clear that the "passive" label carries here no negative connotation.

[11] Since N_+ and N_- will always be finite, an agent's impact on the overall (average) level $\bar{a}e$ of public good or public bad in the network remains negligible. Thus, even if he is also a recipient of it (e.g., pollution, tax compliance) he still takes it as exogenous: his decisions depend only on his role as a *source* of externalities, which he cares about intrinsically.

$e_- < e^* < e_+$. Ex post, there is a *single realization* of the signal, e.g., some salient news or current event, which is then "injected" at random points into the network: each agent i receives it independently with constant probability x. They may also receive it endogenously from their predecessor, and in either case they then choose whether or not to share it with $i+1$, who may or may not also have learnt it directly, can pass it on or not, etc. While abstracting from simultaneous competition between opposing rationales, this framework will nonetheless allow us, by averaging across realizations of e, to analyze how the conversations, beliefs and behaviors of a society and different subgroups within it are shaped by the opposing influences of both types of arguments.

2. Applications. A natural example is that of environmental externalites. Polluters, energy consumers, and litterers impose costs on others, but the magnitude of these costs is subject to dispute and different arguments serve as excuses for inaction or exhortations to change.

A second application is how a dominant group will treat, and justify treating, members of a less powerful one: ethnic minority, immigrants, or women. In the workplace, for instance, members of a majority population engage in behaviors and speech that affect minorities, but they may be unsure of whether those will be experienced as innocuous choices and remarks, offending stereotypes, or traumatizing harassment. Narratives supporting one view or the other circulate, both publicly relayed by the media or court cases about discrimination (probability x) and passed on between people: personal experiences, #metoo testimonies, movies and popular culture, etc. Some majority members genuinely care about not harming or disrespecting minorities, (v_H), others are indifferent or hostile (v_L), but all predominantly want to be seen as benevolent.[12]

Another important case is that of redistribution, whether domestic or toward the developing world. To what extent are the poor really suffering and helpless, and how much good (e) does a charitable contribution or a public transfer (if we interpret a as individual tax compliance, or as voting on the level of public spending, taking its composition as given) really do? Depending on whom one talks to, they will offer arguments and even hard evidence that transfers can make a vital difference to needy people's health and their children's education, improve social cohesion, etc., or that they are often captured by government and NGO bureaucracies, misspent by corrupt local officials, or wasted by recipients themselves on drugs and alcohol. Another narrative of the second kind is that transfers actually harm the poor, by collectively trapping them into a culture of welfare dependency (e.g., Somers and Block 2005).

3. Key tradeoffs. Passive agents' only concern is the behavior of others, so any $i \in P$ will systematically censor antisocial narratives e_- and pass on prosocial ones e_+ when they can make a difference. For $i \in A$, communicating e_- to $i+1$ while choosing $a_i = 0$ has reputational value, but on the other hand it may trigger a *cascade of bad behavior:* inducing the recipient to also act badly (if $i+1 \in A$ and he did not get the signal independently) and furthermore to pass on the excuse to $i+2$, who may then behave in the same way, etc. Conversely, sharing e_+ may induce a *chain of good behavior,* but takes away ignorance as an excuse for one's choosing $a_i = 0$; it also weakens the reputational benefit of choosing $a_i = 1$. Reputation concerns are the same for both moral types but the v_H ones have a stronger influence concern, so they are more

[12] When hurting an outgroup is seen as public good for the ingroup ("keep them in their place, teach them a lesson") only the interpretation of e changes. For multiple audiences with conflicting notions of prosociality, see Appendix B.

inclined to spread positive narratives and refrain from spreading negative ones.[13] The strength of influence motives also depends on how much further an argument is expected to be spread and affect decisions, giving rise to endogenous *social multipliers* that will play a major role in the analysis.

4. Equilibrium. To limit the number of cases, we focus on (stationary) equilibria with the following properties:

(1) *Reputation preservation prevails over influence concerns.* Whenever they learn of a receivable argument for behaving according to self interest, $e = e_-$, both active types choose $a_i = 0$ and invoke the argument that the externality is low, even though this may trigger a chain of bad behavior.[14] In contrast, the influence effect will play a critical role in the propagation of positive narratives, $e = e_+$.

(2) *Common social expectations.* In all instances when they did not learn any narrative, whether directly or from their predecessor, high-type agents (endogenously) choose the same *"default action,"* which we shall denote as $a_H(\emptyset) = 1$ or $a_H(\emptyset) = 0$.[15] This default can be interpreted as the *prevailing social norm,* which can be either prosocial or selfish: it is what a moral person is expected to do (and does), absent any good argument to the contrary. We shall analyze both cases in turn, denoting:

$$x_-^P \equiv \Pr\left[i \text{ knows } e \mid i \in P,\ e = e_-\right], \quad x_-^A \equiv \Pr\left[i \text{ knows } e \mid i \in A,\ e = e_-\right], \tag{7}$$

$$x_+^P \equiv \Pr\left[i \text{ knows } e \mid i \in P,\ e = e_+\right], \quad x_+^A \equiv \Pr\left[i \text{ knows } e \mid i \in A,\ e = e_+\right]. \tag{8}$$

3.2 When acting morally "goes without saying"

Consider first the case where $a_H(\emptyset) = 1$, meaning that high types always behave prosocially *unless* they have an exculpatory narrative. Conversely, observing $a_i = 1$ reveals that they do not have one. When they do learn of e_- (directly or from $i-1$), all active agents choose $a_i = 0$ and pass on the excuse, since (as explained above) we focus on the case where reputational concerns dominate influence ones; the reputation following such disclosure is then $v_D = \bar{v}$. Responsibility narratives e_+, on the other hand, are passed on by no one (active or passive), given any small disclosure cost. Indeed, they do not change any behavior down the line since $a_H(\emptyset) = 1$ already, and on the reputational side they would be redundant for the high type (as $a_i = 1$ is fully revealing) and self-incriminating for the low type. Making use of (6), it follows

[13] The stark distinction between active and passive agents is made for simplicity. Qualitatively similar results would obtain if P's also acted but with low enough observability or reputational concern.

[14] In relatively small groups, a person may sometimes forgo using an available excuse and just "take the blame" for behaving badly so as not to risk enabling similar actions by others: parent in front of their children, leader seeking to instill a "no excuses" culture in an organization, etc. Those cases are relatively rare when agents are small relative to network size, as in our model and the main applications discussed above.

[15] Depending on whether i's "silent" predecessor was a P or an A, and in the latter case on his a_{i-1}, agent i's inferences about e will differ, as we shall see. By restricting attention to equilibria in which i takes the same action across these contingencies, we are thus abstracting from potential others, in which responses differ. This selection limits the number of cases to consider, and the main insights and tradeoffs do not depend on it.

that

$$x_-^P = x + (1-x)(1-\lambda)x_-^A, \quad x_-^A = x + (1-x)\lambda x_-^A, \tag{9}$$

$$x_+^P = x_+^A = x. \tag{10}$$

Consider next agents' inferences when their predecessor does not offer any narrative.

Case 1. Predecessor is an active agent. (a) If $i-1$ chose $a_{i-1} = 0$ without providing an excuse, he must be a low type (as high ones only choose $a = 0$ when they have one available) and either $e = e_-$ but he did not know it (or else he would have disclosed), or $e = e_+$, in which case he does not disclose it even when he knows. Agent i's posteriors over v and e are thus:

$$\hat{v}_{ND} \equiv E\left[v|a_{i-1} = 0, ND\right] = v_L, \tag{11}$$

$$\hat{e}_{ND} \equiv E\left[e|a_{i-1} = 0, ND\right] = \frac{f_-(1-x_-^A)e_- + f_+e_+}{f_-(1-x_-^A) + f_+} > e_0 \tag{12}$$

where ND stands for "no disclosure."

(b) If $i-1$ chose $a_{i-1} = 1$ he must be a high type, and either $e = e_-$ but he did not know it (otherwise he would have chosen $a_{i-1} = 0$ and disclosed) or else $e = e_+$, in which case he does not disclose, since such signals have neither valuable reputational benefits (given $a_{i-1} = 1$) nor influence on anyone's action (as $a_H(\emptyset) = 1$). Therefore, upon observing $(a_{i-1} = 1, ND)$, the updated reputation for $i-1$ is v_H, but the inferences concerning e are again $\hat{e}_{ND}$.

Case 2. Predecessor is passive. When $i-1 \in P$, lack of disclosure conveys no information, since such agents pass on neither e_- (socially harmful) nor e_+ (superfluous given the prevailing norm). The posterior about e thus remains equal to the prior, e_0.

Consider now the tradeoffs involved in the decisions a_i of active types. We shall denote by N_-^A and N_+^A the expected influences that an *active* agent's passing on a narrative e_- or e_+, respectively, have on all of his successors' cumulated contributions. Given the conjectured equilibrium strategies, $N_+^A = 0$: passing on e_+ to a successor has no impact and will thus never be chosen, given an arbitrarily small cost of disclosure. Sharing e_-, on the other hand, will have influence if $i+1$ did not already know of it and happens to also be an active agent (as passive ones take no action and transmit no excuses). More specifically, if he is a high type he will also switch from the default $a_H(\emptyset) = 1$ to $a_{i+1} = 0$ and pass on the excuse. If he is a low type he would have chosen $a_{i+1} = 0$ anyway, but will now also invoke and transmit the excuse, thus influencing followers' behaviors to an extent measured again by N_-^A. Thus:

$$N_-^A = (1-x)\lambda(\rho + N_-^A) \iff N_-^A = \frac{(1-x)\lambda\rho}{1-(1-x)\lambda}. \tag{13}$$

The full set of conditions for an equilibrium with $a_H(\emptyset) = 1$ is thus:

$$v_H e_- N_-^A \leq \mu(\bar{v} - v_L), \tag{14}$$

$$v_H e_-(1 + N_-^A) - c \leq \mu(\bar{v} - v_H), \tag{15}$$

$$v_H \hat{e}_{ND} - c > \mu\left(v_L - v_H\right), \tag{16}$$

The first one states that, when informed of e_-, even a high type will disclose it and choose $a = 0$, rather than doing so without disclosure: the negative social impact is less than the reputational benefit, which is to earn $\bar{v}$ following such action-disclosure pairs rather than v_L for those who behave antisocially without an excuse. The second condition states that he also does not want to choose $a_i = 1$ and censor the news that $e = e_-$. Both inequalities show that disclosures of negative narratives are *strategic substitutes*, in that a higher propensity N_-^A of successors to repeat them makes one more reluctant to invoke them.

The third condition, finally, states that a high active type who received neither a private signal nor a narrative from his predecessor indeed prefers to choose $a_i = 1$ and reveal himself rather than $a = 0$, which given that he has no available excuse would misidentify him as a low type.[16] Finally, together with (12), the condition shows that an equilibrium with the norm $a_H(\emptyset) = 1$ requires that the prior f_+ be high enough, which is quite intuitive.

Proposition 1 (morality as the default behavior). *When (11) -(12) and (14)-(16) hold, there is an equilibrium in which the default (uninformed) action of high types is $a_H(\emptyset) = 1$ and:*

1. *Positive narratives or responsibilities, e_+, are transmitted by no one, since they do not change behavior ($N_+^A = N_+^P = 0$).*

2. *Negative narratives or excuses e_- are transmitted by all active agents, both high- and low-morality.*

3. *The social impact of sharing an excuse is $-e_- N_-^A$, where the virality factor N_-^A is given by (13); such disclosures are therefore strategic substitutes.*

4. *Greater mixing between active and passive agents (lower λ) reduces the multiplier, which both expands the range of parameters for which an equilibrium with moral default action exists, and raises the aggregate provision of the public good or externality within it:*

$$\bar{e} = \frac{\rho}{2}\left(f_+ e_+ + f_-(1 - x_-^A)e_-\right).$$

The intuition for the last and key result is simple. Behavior of the (high) active types departs from the default moral action only when they learn of e_-; since such news are transmitted by both active types and censored by passive types, such learning occurs more frequently, the greater the probability λ that an active agent i is preceded by another active one; similarly, it will travel further, the more likely it is that $i + 1$ is also active.[17]

[16] This requirement corresponds to the more stringent case where the "silent" predecessor is a passive agent, since we saw that nondisclosure by $i \in P$ leads to lower beliefs about e than when $i - 1 \in A$: that is why the expected externality involved is $\tilde{e}_{ND} < e_0$ rather than $\hat{e}_{ND} > e_0$.

[17] One can also show (see the Appendix) that a lower x also raises $\bar{e}$. The lower probability that any active, high-type agent will learn of e_- and pass it on dominates the countervailing effect that his disclosure is more likely to be new information for his successors.

3.3 When "silence is complicity"

Consider now the case where $a_H(\emptyset) = 0$, so that high types behave socially *only* in the presence of a responsibility narrative, $e_+ > e^*$. This, in turn, makes positive-influence concerns relevant for everyone. In particular, a v_H active agent i who knows e_+ will now pass it on to $i+1$, even though $a_i = 1$ already reveals all there is to know about v_i and e. The reason he does so is that $i+1$, or/and some $i+k$ down the line from him, could turn out to be an uninformed passive agent and thus unable to signal e_+ through his actions. Being given the actual narrative will allow $i+1$ to relay it to $i+2$, who may then behave better (if he is a high-type active agent who did not directly learn of e) and/or pass it on to $i+3$ (if he is either a high type or another inactive agent), and so on.

A low type, on the other hand, faces a tradeoff: by sharing e_+ he induces good behaviors among others, but also forsakes the "cover" of pleading ignorance for his own choice of $a_i = 0$. We shall find conditions such that the low type prefers pooling with the uninformed high types, and thus again *censors* positive narratives e_+. As before, both active types pass on negative ones, e_-. Given these action and communication strategies,

$$x_-^P = x + (1-x)(1-\lambda)x_-^A, \quad x_-^A = x + (1-x)\lambda x_-^A, \tag{17}$$

$$x_+^P = x + (1-x)\left[\lambda x_+^P + (1-\lambda)\rho x_+^A\right], \quad x_+^A \equiv x + (1-x)\left[(1-\lambda)x_+^P + \lambda\rho x_+^A\right], \tag{18}$$

where the last two equations reflect the fact that if $i-1 \in A$ and knows that $e = e_+$ he discloses it when he is a high type. Thus x_-^P and x_-^A are unchanged from the previous case, but x_+^P and x_+^A are more complicated; see (A.1)-(A.2) in the Appendix. The "influence factors" or social multipliers are now $N_-^A = N_-^P = 0$ for e_- (as it will change no behavior), while for e_+ they are

$$N_+^P = (1-x)\left[\lambda N_+^P + (1-\lambda)\rho(1+N_+^A)\right], \tag{19}$$

$$N_+^A = (1-x)\left[\lambda\rho(1+N_+^A) + (1-\lambda)N_+^P\right], \tag{20}$$

for passive and active agents (of either moral type), respectively. The solutions to this linear system are given by (A.4)-(A.5) in the Appendix.

Consider now the updating.

(a) As before, any active agent who chooses $a_{i-1} = 0$ but provides an excuse e_- receives the pooling reputation $\hat{v}_D = \bar{v}$. For those who do not have one, however, the equilibrium is now more "forgiving":

$$\hat{v}_{ND} = \frac{\rho(1-\bar{x}^A)v_H + (1-\rho)[1-f_-x_-^A]v_L}{\rho(1-\bar{x}^A) + (1-\rho)\left[1-f_-x_-^A\right]} \in (v_L, \bar{v}), \tag{21}$$

where $\bar{x}^A \equiv f_+x_+^A + f_-x_-^A$. Indeed, $i-1$ could be a high type who was uninformed (probability $1-\bar{x}^A$), as well as a low type who either was uninformed or received but censored e_+ (total probability $1-f_-x_-^A$). As to the expected externality following such an observation, it is

$$\hat{e}_{ND} \equiv E\left[e \mid a_{i-1} = 0, ND\right] = \frac{f_-(1-x_-^A)]e_- + f_+(1-\rho x_+^A)e_+}{f_-(1-x_-^A) + f_+(1-\rho x_+^A)} > e_0. \tag{22}$$

(b) If the "silent" predecessor $i-1$ was a passive agent, on the other hand, he will pass on e_+ but censor e_-, so i's inference about e is

$$\tilde{e}_{ND} \equiv E\left[e \mid i-1 \in P, ND\right] = \frac{f_- e_- + f_+(1-x)e_+}{f_- + f_+(1-x)} < e_0. \tag{23}$$

Lack of disclosure by actors is thus *positive* news about e since their dominant concern is preserving reputation, whereas lack of disclosure by passive agents is *negative* news about e since their sole concern is minimizing others' misbehavior; formally, $\hat{e}_{ND} > e_0 > \tilde{e}_{ND}$.

The conditions for an equilibrium with the norm $a_H(\emptyset) = 0$ are then

$$v_L e_+ N_+^A < \mu\left(\hat{v}_{ND} - v_L\right), \tag{24}$$

$$c - v_H \hat{e}_{ND} \geq \mu(v_H - \hat{v}_{ND}), \tag{25}$$

where $\hat{v}_{ND}$ is defined by (21).[18] Condition (24) states that, when learning e_+, a low type agent prefers to *keep quiet* about it and maintain the pooling reputation $\hat{v}_{ND}$ rather than reveal himself, even though this information retention will prevent on average N_+^A (high-type) followers from switching to the prosocial action. The inequality also demonstrates that for positive narratives, sharing decisions are *strategic complements.* The more others tend to pass them on (higher N_+^A), the greater is the (now positive) externality that will result from i's revealing such a signal; consequently, the higher the "self-incrimination" concern must be to prevent him from essentially communicating: "do as I say, not as I do."

Condition (25) states that, absent any narrative, a high type indeed chooses $a_H(\emptyset) = 0$ rather than deviating to $a_i = 1$, which would clearly identify him but not persuade $i+1$ that $e = e_+$, since if he knew that he should have disclosed it.[19] Referring to (22), finally, shows that an $a_H(\emptyset) = 0$ equilibrium requires that the prior f^+ not be too high, which is again intuitive.

Proposition 2 (**selfishness as the default behavior**). *When (21)-(22) and (24)- (25) hold, there is an equilibrium in which the default (uninformed) action of high types is $a_H(\emptyset) = 0$ and:*

1. *Negative narratives or excuses e_- are transmitted by all active agents, both high- and low-morality, but this has no impact on others' behavior ($N_-^A = 0$).*

2. *Positive narratives or responsibilities e_+ are transmitted by both passive agents and high-morality active ones, but concealed by those with low morality.*

3. *The social impact of sharing a positive narrative is $e_+ N_+^A$ for an active agent and $e_+ N_+^P$ for a passive one, where the virality factors N_+^A and N_+^P are given by (A.4) and (A.5).*

[18] Two other conditions automatically hold: (i) $v_H e_- N_-^A = 0 < \mu(\bar{v} - \hat{v}_{ND})$, so active agents will always share an excuse e_-(and choose $a = 0$), as it is reputationally valuable and has no spillover onto followers' behavior; (ii) $c - v_H e_+(1 + N_+^A) < \mu(v_H - \hat{v}_{ND})$, which follows from $e_+ > e^*$ and $\hat{v}_D < \bar{v}$ and embodies the same complementarity as (24): high types learning e_+ have even stronger reasons to choose $a = 1$ (and now disclose it) than in the basic model; in fact, we see that such an equilibrium could even be sustained with $e_+ < e^*$.

[19] In contrast to the previous ($a_H(\emptyset) = 1$) type of equilibrium, the expected externality is now $\hat{e}_{ND} > e_0$ rather than $\tilde{e}_{ND} < e_0$ (with $\hat{e}_{ND}$ given by (22)), namely the belief when i's predecessor was active and chose $a_{i-1} = 0$ –making his silence a signal that e is more likely to be high (whereas if he was passive it would indicate that e is more likely to be low).

Such disclosures are therefore strategic complements.

4. *Greater mixing between active and passive agents (lower λ) lowers N_+^A and raises N_+^P. It both expands the range of parameters for which an equilibrium with immoral default action exists and raises the aggregate provision of the public good or externality within it:*

$$\bar{e} = \frac{\rho}{2} f_+ e_+ x_+^A.$$

The intuition for the last result is that behavior of the (high) active types departs from the default immoral action only when they learn of e_+; such news are transmitted by all passive types, but by only a fraction ρ of active ones. Therefore, an active agent i is more likely to learn of it if his predecessor $i-1$ is passive and, similarly, once he transmits it to $i+1$ it is likely to travel further if $i+1$ is also a passive agent.[20]

3.4 Implications: firewalls, relays and polarization

Note first that the two types of equilibria and social norms are associated with very different circulating narratives. In the "moral" equilibrium ($a_H(\emptyset) = 1$), doing the right thing (e.g., treating fairly the minority or less powerful group) *"goes without saying,"* while deviating requires a justification, so negative narratives are the ones that will get passed on (when they occur) and affect behavior. In the "amoral" equilibrium ($a_H(\emptyset) = 0$) self-interest is the default, but excuses remain valuable and thus again circulate. Now, however, so will positive narratives, propagated by passive and high-morality active agents to push others to behave well; in contrast, *"silence is complicity."*[21]

Second, even though the two types of equilibria involve radically different norms and outcomes, Propositions 1-2 show that in *either* case, more *mixed interactions* (lower λ) *raise prosocial behavior.* Intuitively, agents whose actions and/or morality are not "in question" (irrelevant or unobservable) have no need for excuses, and thus act both as *"firewalls"* limiting the diffusion of exonerating narratives, and as *"relays"* for responsibilizing ones. The latter, furthermore, encourages high-morality actors to do the same, via strategic complementarity. Third, intermingling agents with different stakes in reputation preservation versus social influence leads to a social discourse and set of beliefs that are not only more moral (or "moralizing") *on average*, but also *less polarized,* as we show below.

Proposition 3 (**polarization**). *In either type of equilibrium, the gaps between active and passive agents' awareness of narratives, measured respectively by $\left|\ln(x_-^P/x_-^A)\right|$ for negative ones and $\left|\ln(x_+^P/x_+^A)\right|$ for positive ones, are both U-shaped in the degree of network segregation λ, with a minimum of zero at $\lambda = 1/2$ and a global maximum at $\lambda = 1$.*[22]

When majority and minority individuals (say) interact mostly within segregated pools (high λ), opposite types of narratives will circulate within each one, with the majority group mostly

[20] Here again, a lower x increases (both) multipliers, but it now reduces $\bar{e}$; see the Appendix.

[21] Both equilibria may coexist for some range of parameters (e.g., prior distribution F), with Pareto-dominance having little bite for selection.

[22] For the equilibrium with $a_H(\emptyset) = 1$, one of the U-shapes is degenerate, in that $\ln(x_+^P/x_+^A) = 0$ for all λ. All other statements in Proposition 3 hold in the strict sense, including for the global maximum at $\lambda = 1$.

sharing rationalizations for their behavior (e_-), which will be worse on average than under integration, and the minority group mostly sharing reasons for why it is *inexcusable* (e_+).

Notably, the *polarization-minimizing* pattern differs from the *prosociality-maximizing* one: it is not a deterministic alternation of agents ($\lambda = 0$) but a *random* one ($\lambda = 1/2$), or equivalently, a tree in which each individual's audience is a 50-50 mix of A's and P's. For $\lambda \approx 0$ each A hears only from a P, so he can learn e_- only exogenously, whereas each P hears (only) from an A, so she can also learn it from him. Beliefs again diverge, but it is now minorities who are more likely to hear, from a majority individual, of an excuse for his controversial behavior. Similarly, when responsibility narratives circulate, if $\lambda \approx 0$, majority members are more exposed to them, since any minority individual who knows e_+ will relay it, whereas only high-morality majority individuals will disclose it.

Implications. Whether for ethnicity, gender, or income, positive correlation ($\lambda \geq 1/2$) is by far the most relevant scenario, leading to differences in exposure to exculpatory versus responsibility arguments, and therefore in beliefs, that are of the intuitive rather than the "paradoxical" type. Assortative social communication can arise from reasons taken here as exogenous (e.g, homophily), but also one emanating from the model: whereas the P's want to be "heard" by the A's, the latter have an incentive to "listen" instead to other A's, who are more likely to provide them with excuses and less with responsibility arguments.

4 Moral Standards: What Excuses Are Admissible?

We now turn to the production side of narratives, studying the case where the signal e arises from the agents' own *search for reasons* to act or not to act morally.

Looking for arguments generally serves three purposes: they can help the individual figure out the consequences of his actions (*decision value*), justify them to others or to himself (*reputation value*), and/or convince others to act in certain ways (*influence value*). We shall focus here on the interplay of the first two, which raises new questions due to the fact that high and low-morality types will search differently, so that the mere fact of having an excuse to offer is informative. *How strong* must an excuse then be in order to be socially acceptable? And *how much stigma* is incurred by someone who behaves selfishly without having one, or only one that is too weak?

Absent influence motives, we can focus on a single actor-audience (A, P) pair to analyze these issues. Suppose that, prior to acting but after learning his type, the agent can learn or come up with the narrative $\sigma = e \sim F(e)$ with any probability x, at cost $\psi(x)$. For instance, they can look for information on climate change and its effects. In this section, F is taken here to have full support on $[0, 1]$, so that we can examine excuses of varying strength; with probability $1-x$, he learns nothing, $\sigma = \emptyset$. We assume $\psi(0) = \psi'(0) = 0$, $\psi' > 0$, $\psi'' > 0$ and $\psi(1) = +\infty$, and denote by x_H and x_L the two types' search strategies. When knowing e the agent can disclose it to his audience (or rehearse it for himself), at some infinitesimal cost to break indifferences.[23]

To build up intuition, note that v_H types are very concerned about taking the appropriate

[23] In an intrapersonal context, the search for absolving narratives can also be interpreted as a form of motivated moral reasoning. On that topic, see, e.g., Ditto et al. 2009 and Bénabou and Tirole (2016).

action, so their search intensity x_H will reflect the *option value(s)* of finding out whether e might be especially high or low –that is, the extent to which the distribution $F(e)$ has significant mass in the upper or lower tail. In addition, they also value the fact that, when learning that e is low, disclosing it will reduce the reputational cost of self-interested behavior. Image concerns thus factor into the v_H types's search decisions, but less so than for v_L types, whole *sole interest* in searching is to find low enough values of e to justify behaving selfishly (recall that $a_L \equiv 0$ in the equilibria we focus on). The meaning of merely having an excuse, and thus the acceptability threshold $\hat{e}$, will thus depend on the balance between the *tail risks* of incorrect decisions and the strength of *image concerns.*

It will be useful to define, for any distribution $F(e)$, the two conditional moments

$$\mathcal{M}^-(e) \equiv E_F\left[\tilde{e} \mid \tilde{e} \leq e\right] \quad \text{and} \quad \mathcal{M}^+(e) \equiv E_F\left[\tilde{e} \mid \tilde{e} > e\right], \tag{26}$$

which will govern the option values discussed above, and are linked by the constraint that $F(e)\mathcal{M}^-(e) + [1 - F(e)]\mathcal{M}^+(e) = E_F[e]$ must give back the prior, e_0.

We shall now analyze, proceeding backwards: (a) the inferences made by an audience observing the action, accompanied by disclosure (D) of a narrative e, or by no disclosure (ND); (b) the incentives of an agent who knows of e to disclose it, or say nothing; (c) the incentives to engage in costly search to find out the value of e.

Moral standards. We shall focus attention on equilibria taking the following intuitive form: when the signal e about the importance of the externality is below some cutoff $\hat{e}$, both types disclose this "excuse" and choose $a = 0$. When it is above, the high type chooses $a = 1$, perfectly separating himself, and neither type discloses e, as this would be useless for the high type, and self-incriminating for the low one –such a weak justification would not be accepted.

The common disclosure strategy implies that all equilibrium narratives $e \leq \hat{e}$ have the same informational content about the agent's type: when $a = 0$ is accompanied by such an excuse, the resulting expectation about his type is

$$\hat{v}_D = \frac{\rho x_H v_H + (1-\rho) x_L v_L}{\rho x_H + (1-\rho) x_L}, \tag{27}$$

which is independent of e.[24] Note also that $\hat{v}_D$ is increasing in x_H/x_L, as this implies a higher likelihood that the person who found the excuse is a high type.

The threshold where the high type, when informed, is indifferent between the strategies $(a = 0, D)$ and $(a = 1, ND)$ is then uniquely given by:

$$v_H \hat{e} - c + \mu(v_H - \hat{v}_D) \equiv 0. \tag{28}$$

Note that $\hat{e} > e^*$ when $\hat{v}_D > \bar{v}$, or equivalently $x_L < x_H$, and vice versa. We shall denote as

[24]The denominator is always well-defined, as there is no equilibrium (in undominated strategies) in which $(x_H, x_L) = (0, 0)$; see the "Proofs" section of this Appendix. Note also that, under the self-signaling interpretation in which disclosure of reasons is "to oneself" (e.g., rehearsal), $\hat{v}_D$ and $\hat{v}_{ND}$ further below depends onl on the equilibrium values of (x_L, x_H), and not on the actual (potentially deviating from equilibrium) choice of x. In other words, the individual later on forgets the chosen search intensity x and thus assesses his excuses just as an outside observer would.

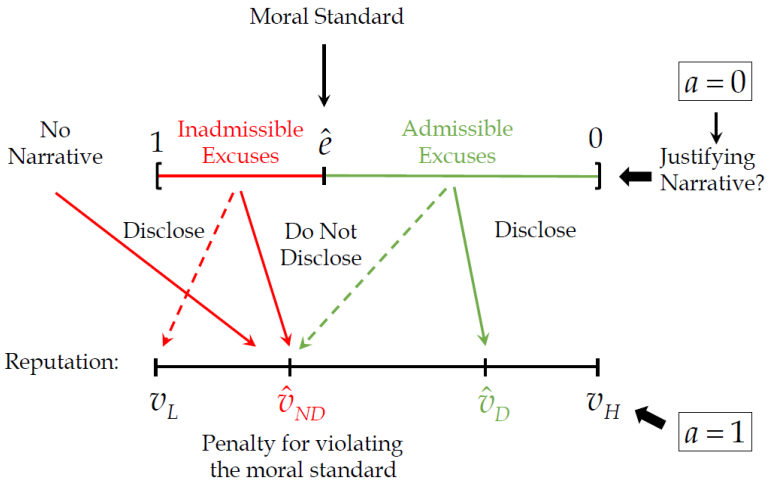

Figure 2: Moral Standards and Narratives. Straight arrows describe equilibrium play at the disclosure stage, dashed ones off-path deviations

$\hat{v}_{ND}$ the audience's posterior when it observes $a = 0$ without a justifying argument. Its value will depend in particular on whether the high type's "default" action –his behavior absent any information– is $a = 1$ or $a = 0$, but it must always be that $\hat{v}_{ND} < \hat{v}_D$.[25]

Intuitively, and as illustrated in Figure 2, $\hat{e}$ and $\hat{v}_{ND}$ define *society's moral standard*, and the penalty for violating it*: how strong* an excuse must be in order to be effective (e must be below $\hat{e}$), and how much *stigma* is incurred for failing to produce one when behaving selfishly ($\bar{v} - \hat{v}_{ND}$). No one offers a "lame" excuse $e \geq \hat{e}$, as this would be reputationally worse than offering no justification (stigma $\bar{v} - v_L$). Taking the case of symmetric-information about e as a benchmark, we will say that the standard is *strict* if $\hat{e} < e^*$ and *lenient* if $\hat{e} > e^*$. Note from (28) that $\hat{e}$ also defines the meaning of having an (acceptable) excuse, namely the inferences $\hat{v}_D$ made when somebody produces one.

While this form of threshold equilibrium is very natural, there could in general be more complicated ones as well, sustained by off-path beliefs that "punish" the disclosure of any arbitrary set N of values of e by attaching to them very low beliefs, such as v_L. Facing such a significant reputation loss, the high type may prefer to choose $a = 1$ when learning $e \in N$, so that not only disclosure but even the choice of a is no longer a cutoff rule. In the Supplementary Appendix we show that imposing a plausible restriction on off-path beliefs eliminates all such equilibria, leaving only the single-threshold class described above.

We next solve for equilibrium behavior, first given any available narrative e, then at the initial stage of searching for narratives.

[25] Otherwise there would be zero disclosure, hence $x_L = 0$, $\hat{v}_D = v_H > \hat{v}_{ND}$ and a contradiction, as long as $x_H > 0$ –and indeed some information is always useful for the high type since $F(e)$ has full support. As to an equilibrium where $x_L = 0 < x_H$ but the high type does not disclose some $e < \hat{e}$ for fear of earning a low reputation, it is ruled out by elimination of strictly dominated strategies; see the Supplementary Appendix B.

4.1 Prosocial norm and finding reasons not to act

1. Action and disclosure. When the prior e_0 is high enough, the high type will choose $a_H(\emptyset) = 1$ when uninformed, so narratives can only provide potential reasons to act *less* morally. In such an equilibrium, when the audience observes $a = 0$ without an excuse it knows that the agent is a low type, so $\hat{v}_{ND} = v_L$. The high type will then indeed act morally unless there is a good reason *not* to, that is, as long as $v_H e_0 - c + \mu(v_H - v_L) \geq 0$, or substituting in (3):

$$v_H(e_0 - e^*) \geq \mu(v_L - \bar{v}) = -\mu\rho(v_H - v_L). \tag{29}$$

As expected, this defines a minimal value for e_0, which is below e^* since the right-hand side is negative. When learning the value of e, on the other hand, it is optimal for the high type to choose $a = 1$ (and not waste the small disclosure cost) if $e > \hat{e}$ given by (28), whereas if $e \leq \hat{e}$ it is optimal to disclose it (since $\hat{v}_D > \hat{v}_{ND}$) and choose $a = 0$.

2. Search. Consider now the optimal search strategy of the high type. If he learns that the state is $e < \hat{e}$, he will disclose it and choose $a = 0$, leading to a utility of $\mu\hat{v}_D$. If he does not have such an excuse, having either not looked for one, failed in his search ($\sigma = \emptyset$) or found out that $e \geq \hat{e}$, he will choose $a = 1$, and achieve $v_H e - c + \mu v_H$. His expected utility from a search intensity x is therefore

$$\begin{aligned} U_H(x) = -\psi(x) + x\left[\mu F(\hat{e})\hat{v}_D + \int_{\hat{e}}^{1} (v_H e - c + \mu v_H) dF(e)\right] \\ + (1 - x)\int_0^1 (v_H e - c + \mu v_H) dF(e), \end{aligned}$$

leading to the first-order condition

$$\psi'(x_H) = F(\hat{e})\left[c - \mu(v_H - \hat{v}_D) - v_H \mathcal{M}^-(\hat{e})\right] = F(\hat{e}) v_H \left[\hat{e} - \mathcal{M}^-(\hat{e})\right]. \tag{30}$$

The low type, trying to mimic the high one, will only disclose those same values $e \leq \hat{e}$, when he knows them. When no excuse is available ($\sigma = \emptyset$), on the other hand, his action reveals that he cannot be the high type, who chooses $a = 1$ unless a good reason not to can be provided. The low type's ex-ante utility from searching with intensity x is thus

$$U_L(x) = -\psi(x) + xF(\hat{e})\mu\hat{v}_D + \left[1 - xF(\hat{e})\right]\mu v_L,$$

leading to

$$\psi'(x_L) = \mu F(\hat{e})(\hat{v}_D - v_L). \tag{31}$$

3. Equilibrium. An equilibrium is a quadruplet $(x_H, x_L, \hat{e}, \hat{v}_D) \in [0,1]^3 \times [v_L, v_H]$ satisfying equations (27)-(28) and (30)-(31), together with a prior e_0 high enough that inequality (29) holds. Furthermore, $x_H > x_L$ if and only if $\mathcal{M}^-(\hat{e}) v_H \leq c - \mu(v_H - v_L)$ or, equivalently

$$\mathcal{M}^-(\hat{e}) v_H \leq c - \mu(v_H - v_L). \tag{32}$$

Intuitively, the high type is more eager to learn e when there is a substantial probability

that it could be very low, as this has high decision-making value. Thus (30) shows that x_H rises, ceteris paribus, as $\mathcal{M}^-(\hat{e})$ declines and/or $F(\hat{e})$ rises. The low type, in contrast, is interested in narratives only for their exculpatory value, which does not depend on e as long as it is low enough that the high type would also invoke it. Comparisons of tail moments will thus play an important role, so we define:

Definition 1. *Given a cutoff $e^\dagger \in (0,1)$, a distribution F_1 is more $e^\dagger$-bottom heavy than another distribution F_2 if $\mathcal{M}^-_{F_1}(e^\dagger) < \mathcal{M}^-_{F_2}(e^\dagger)$. Conversely, F_1 is more $e^\dagger$-top heavy than F_2 if $\mathcal{M}^+_{F_1}(e^\dagger) > \mathcal{M}^+_{F_2}(e^\dagger)$. If F_1 and F_2 have the same mean and $F_1(e^\dagger) = F_2(e^\dagger)$, these two properties are equivalent.*

The following lemma provides two sufficient conditions relating this property to familiar ones. The first one allows F_1 and F_2 to have the same mean (F_1 is then a mean-preserving spread of F_2), while the second precludes it.

Lemma 1.

1. *Let F_1 be second-order stochastically dominated by F_2. If $F_1(e^\dagger) \leq F_2(e^\dagger)$, then F_1 is more $e^\dagger$-bottom heavy than F_2; if $F_1(e^\dagger) \geq F_2(e^\dagger)$, then F_1 is more $e^\dagger$-top heavy than F_2.*

2. *If the likelihood ratio f_2/f_1, or more generally, F_2/F_1, is increasing, then F_1 is more $e^\dagger$-bottom heavy than F_2 at all $e^\dagger$. If f_1/f_2, or more generally, $(1-F_1)/(1-F_2)$, is increasing, then F_1 is more $e^\dagger$-top heavy than F_2 at all $e^\dagger$.*

Formalizing the previous intuitions about each type's incentive to search for excuses, we can now state the following results.

Proposition 4 (prosocial norm). *For any e_0 high enough that (29) holds, there exists an equilibrium where moral behavior is the default (uninformed) choice of the high type ($a_H(\emptyset) = 1$) and violating the moral standard (behaving selfishly without a narrative $e \leq \hat{e}$) carries maximal stigma ($\hat{v}_{ND} = v_L$). In any such equilibrium, moreover:*

1. *If the distribution of signals $F(e)$ is sufficiently e^*-bottom heavy, in the sense that*

$$v_H[e^* - \mathcal{M}^-(e^*)] > \mu\rho(v_H - v_L), \tag{33}$$

   the high type is more likely to search for narratives: $x_H > x_L$, and correspondingly producing an admissible excuse ($e < \hat{e}$) when choosing $a = 0$ improves reputation, $\hat{v}_D > \bar{v}$. The potential existence of many strong reasons for not taking the moral action (bottom-heaviness of F) makes coming up with even a relatively weak one less suspect, which in turn makes the moral standard more lenient than under symmetric information ($\hat{e} > e^*$).

2. *If $F(e)$ is sufficiently e^*-bottom light that (33) is reversed, it is the low type who is more likely to search for narratives: $x_H < x_L$, so producing an admissible excuse ($e < \hat{e}$) worsens reputation, $\hat{v}_D < \bar{v}$. The fact that most reasons for not taking the moral action one could hope to find are relatively weak ones (top-heaviness of F) implies that coming up with even a strong one raises suspicions about motives, which in turn makes the moral standard more strict than under symmetric information ($\hat{e} < e^*$).*

Intuitively, $e^* - \mathcal{M}^-(e^*)$ scales the high type's option value of finding out whether e may be low enough that, under perfect information, he would prefer $a = 0$. It is thus naturally larger, the worse is the conditional mean of e below e^*, corresponding to bottom-heaviness. The term on the right of (33), on the other hand, is the reputational value of having an excuse available when choosing $a = 0$, which is equally valuable for both types. These observations lead to further comparative-statics results.

Proposition 5. *Let $F(e)$ have the monotone-hazard-rate property. As the reputational incentive $\mu(v_H - v_L)$ rises due to a change in any of its components, condition (33) becomes less likely to hold, making the equilibrium more likely to be of the type where $x_H < x_L$ and the moral standard is strict ($\hat{e} < e^*$).*

4.2 Selfish norm and finding reasons to act

When the prior e_0 is low, intuition suggests that the high type will choose $a_H(\emptyset) = 0$ when uninformed. Narratives can now only provide potential reasons to act *more* morally, and this is the "good" reason why the high type searches for them. post, of course, the signal may turn out to be low, justifying inaction, and that is why the low type searches for them as well.

1. Action and disclosure. In equilibrium, both types reveal all values of $e \leq \hat{e}$ (when they know them), resulting in reputation $\hat{v}_D$ still given by (27) and the same threshold $\hat{e}$ as in (28). Beliefs following $(a = 0, ND)$, however, are now

$$\hat{v}_{ND} = \frac{\rho\,(1 - x_H)\,v_H + (1 - \rho)[1 - x_L F(\hat{e})]v_L}{\rho\,(1 - x_H) + (1 - \rho)[1 - x_L F(\hat{e})]} > v_L, \tag{34}$$

which is decreasing in x_H and decreasing in $x_L F(\hat{e})$. A selfish action without an accompanying excuse is thus less damaging to reputation than in the previous case, since it may now come from an uninformed high type (recall that, in the type of equilibrium considered, $a_H(e) = 1$ when $e \geq \hat{e}$). When there is an admissible excuse $e < \hat{e}$, conversely, disclosing is indeed optimal. Given these reputational values, the uninformed high type will indeed prefer not to act, $a_H(\emptyset) = 0$, if $v_H e_0 - c + \mu(v_H - \hat{v}_{ND}) \leq 0$ or, equivalently

$$v_H(e_0 - e^*) \leq \mu(\hat{v}_{ND} - \bar{v}). \tag{35}$$

As expected, this now puts an upper bound on the prior e_0 about the severity of the externality ($e_0 v_H \leq c$). Conversely, even though $\hat{v}_{ND}$ depends on the distribution F and thus on its mean e_0, (35) will be shown to hold whenever e_0 is low enough.

2. Search. Computing again the expected utilities $U_H(x)$ and $U_L(x)$ now leads to the optimality conditions (see the Appendix):

$$\psi'(x_H) = \mu\,(\hat{v}_D - \hat{v}_{ND}) + [1 - F(\hat{e})]\,[\mathcal{M}^+(\hat{e}) - \hat{e}]v_H, \tag{36}$$

$$\psi'(x_L) = \mu F(\hat{e})\,(\hat{v}_D - \hat{v}_{ND})\,. \tag{37}$$

so it must always be that $x_H > x_L$, which as noted earlier implies that $\hat{v}_D > \bar{v}$ and $\hat{e} > e^*$.[26]

3. Equilibrium. This is now a quintuplet $(x_H, x_L, \hat{e}, \hat{v}_D, \hat{v}_{ND}) \in [0,1]^3 \times [v_L, v_H]^2$ satisfying equations (27)-(28), (34) and (36)-(37), together with a prior e_0 low enough for inequality (35) to hold. The basic intuition shaping the equilibrium is that, since the high type is now also interested in finding out about high values of e (as these will switch his decision to $a_H = 1$), it is now always he who searches more intensively for narratives, compared to the low type.

Proposition 6 (**selfish norm**). *For any e_0 low enough, there exists an equilibrium where abstaining is the default (uninformed) choice of the high type ($a_H(\emptyset) = 0$) and violating the moral standard (behaving selfishly without a narrative $e \leq \hat{e}$) carries only moderate stigma ($\hat{v}_{ND} > v_L$). In any such equilibrium, moreover:*

1. *The high type is more likely to search for narratives, $x_H > x_L$, so if they are disclosed on the equilibrium path (following $a = 0$), producing one improves reputation, $\hat{v}_D > \bar{v} > \hat{v}_{ND}$.*
2. *The high type's strong desire to look for positive narratives makes coming up with even a negative one less suspect, and as a result makes the moral standard more lenient. ($\hat{e} > e^*$).*

Interestingly, equations (29) and (35) can be shown to be compatible over a range of priors, so that both types of equilibria can coexist.

Proposition 7 (**multiple norms and moral standards**). *Let $\psi'(1) = +\infty$. There is a nonempty range $[\underline{e}_0, \bar{e}_0]$ such that, for any prior e_0 in that interval, there exists both:*

(i) A strict-moral-standard equilibrium ($\hat{e} < e^*$), in which the default choice of the high type is to act prosocially ($a_H(\emptyset) = 1$) and reputation suffers when failing to do so even with a good excuse ($\bar{v} > \hat{v}_D > \hat{v}_{ND} = v_L$).

(ii) A lenient-moral-standard equilibrium ($\hat{e} > e^*$), where the default is to act selfishly ($a_H(\emptyset) = 0$) and providing a good excuse for doing so enhances reputation (though less than acting morally: $v_H > \hat{v}_D > \bar{v} > \hat{v}_{ND} > v_L$).

To summarize, we showed that two key factors determe whether a prosocial or selfish culture tends to prevail, and whether standards for excuses are strict or lenient. First, and quite naturally, people's prior mean e_0 about whether individual actions have important or minor externalities. Second, and more subtly, the tail risks in the uncertainty surrounding that question. For instance, keeping e_0 fixed, suppose that people perceive even a small probability that some group could be very "undeserving" –not providing complementary efforts, or even hostile, treacherous, etc. e. Formally, e *could* be very low. That fear will justify "looking into it," and even when such scrutiny reveals only less serious concerns (e.g., isolated cases or anecdotes), lowering e only slightly from e^*), such narratives can be socially acceptable reasons for both types to treat that group badly. There are now "excuses for having excuses," even when the latter are weak ones, and as a result this erodes moral standards.

When multiple norms and associated standards for excuses coexist, the question arises of which one agents want to and/or can coordinate on. From the point of view of a single individual, as before both types tend to prefer operating under a more lenient standard (playing the $a_H(\emptyset) = 0$ equilibrium, when it exists), at least when x_H and x_L are exogenous and equal

[26] Clearly, $x_H \geq x_L$. Equality would mean that $\hat{v}_D = \bar{v}$ and hence $\hat{e} = e^* < 1$, which given full support of f would imply that $F(\hat{e}) < 1$ and $\mathcal{M}^+(\hat{e}) > \hat{e}$; (36) would then lead to $x_H > x_L$, a contradiction.

(corresponding to an extreme form of the function ψ); when they are endogenous, in general one cannot rank the equilibria. From the aggregate, societal point of view, moreover, if each actor is himself subject to the externalities created by many others (e.g., pollution), the more prosocial equilibrium $a_H(\emptyset) = 1$ will tend to be collectively preferred, especially if F is top-heavy (and c not too large).

5 Conclusion

We developed a tractable framework for jointly analyzing moral behavior and moral discourse. Along with standard factors such as intrinsic preferences and image concerns, it brings into play how agents search for, strategically use, and socially circulate arguments and narratives about the moral significance of different actions.

The model could be applied to other issues, such as cultural or political identity. It would also be interesting to extend it to richer network structures than our simple linear one –for instance, in the line of Campbell et al.(2025) but with agents linked through information transmission rather than adoption complementarities. We modeled narratives as acting like hard signals about social or/and private payoffs, while stressing that, in practice, they need not have real informational content. Put differently, we took as a primitive a class of arguments that "work" in persuading agents and focused on analyzing how people will then search for them, invoke them, and judge those who do so. A promising direction for future research would be to combine models that explicitly incorporate heuristic or motivated cognitive processes allowing biased narratives to persuade (e.g., Eliaz and Spiegler 2020, Schwartzstein and Sunderam 2021) with the present framework, in which agents communicate not just through messages but also through observable actions.

Conflicting narratives represent another interesting direction for further work, as in our model a single (but stochastically varying) argument is circulating at any given time. In existing models of narrative competition, politicians or firms costlessy offer rival narratives designed to manipulate voters' or consumers' behavior. Contexts such as morality or identity involve elements of both frameworks: different rationales for what is right or wrong compete –often put forward by norms entrepreneurs– but a person's words must be consistent with their actions and people seek not just to influence others but also to justify their own behavior.

Differing social preferences even under full information or heterogenous priors constitute another potential source of disagreement, often relevant for societal issues such as religion, abortion, immigration, etc. When audiences disagree on what constitutes a negative versus positive externality, social image becomes multidimensional, and the net effects of these concerns for different individuals will reflect the degree of assortative matching in society. The latter is ultimately endogenous, however, so it would be interesting to analyze group or network formation when agents communicate both trough what they say and through what they do.

Appendix

Audiences with incompatible values. We show here how the specification (1) extends to such situations, as claimed in the text. Let there be two groups but still one action, which creates (perceived) externalities $e_1 > 0$ for Group 1 and $-e_2 < 0$ for Group 2, internalized by the agent as $v_1 e_1 - v_2 e_2$, $(v_1, v_2) \in \{v_L, v_H\}^2$. Each group esteems and rewards people who they think "care" about it or "have the right values," i.e. are inclined to actions that it deems beneficial; conversely, it shames and punishes those it perceives as likely to inflict harms.

A first simple specification is where $e_1 = e_2 > 0$, making $v \equiv v_1 - v_2$ a sufficient statistic and $U = ve - c + (\mu_1 - \mu_2)E[v|a]$. Alternatively, let agents care only about one externality, while still valuing reputation in both groups: $v_2 \equiv 0$, so $U = v_1 e_1 - c + (\mu_1 - \mu_2)E[v_1|a]$. In either case, the model is essentially unchanged, provided that $(\mu_1 - \mu_2)(v_1 - v_2) > 0$ for each agent: he cares more about his social standing in the eyes of the group that has values closer to his own (e.g., due to partial assortative matching, or prospects thereof). A more general and symmetric model, but now truly multidimensional and thus more complex, would be one with: (i) three actions, $a = 0, 1, 2$, where actions 1 and 2 favor Groups 1 and 2 respectively and are (equally) costly, whereas the neutral choice 0 (doing nothing) is not; (iii) five types, $(v_1, 0)$ and $(0, v_2)$, with $(v_1, v_2) \in \{0, v_L, v_H\}^2$.

Proof of Proposition 1 Only the last result remains to show. Since 1/2 of agents are active with a fraction ρ of them high types, and each has probability x_-^A (given by (9)) of being informed of e_- when it occurs, we have:

$$\bar{e} = \frac{\rho}{2}\left[f_+ e_+ + f_-(1 - x_-^A)e_-\right] = \frac{\rho}{2}\left[f_+ e_+ + f_- e_- \frac{(1-x)(1-\lambda)}{1-(1-x)\lambda}\right].$$

which is decreasing in λ and in x, assuming $e_- > 0$. ■

Proof of Proposition 2 We first solve the system (18) to obtain

$$x_+^P = [1 - (1-x)\rho(2\lambda - 1)]\left(\frac{x}{Z}\right), \tag{A.1}$$

$$x_+^A = [1 - (1-x)(2\lambda - 1)]\left(\frac{x}{Z}\right). \tag{A.2}$$

where

$$\begin{aligned} Z &\equiv [1-(1-x)\lambda][(1-(1-x)\rho\lambda] - (1-x)^2(1-\lambda)^2\rho \\ &= 1 - (1-x)(1+\rho)\lambda + (1-x)^2\rho(2\lambda - 1) \end{aligned} \tag{A.3}$$

Turning next to system in N_A^+ and N_P^+, it yields

$$N_A^+ = \frac{(1-x)\rho(\lambda - (2\lambda - 1)(1-x))}{Z} \tag{A.4}$$

$$N_P^+ = \frac{(1-x)(1-\lambda)\rho}{Z}. \tag{A.5}$$

To show that $\partial N_+^A/\partial\lambda > 0$, we compute the determinant,

$$\begin{vmatrix} 2x-1 & 1-x \\ 2\rho(1-x)^2-(1+\rho)(1-x) & 1-\rho(1-x)^2 \end{vmatrix}$$
$$= (2x-1)(1-\rho(1-x)^2) - (1-x)^2[2\rho(1-x)-(1+\rho)]$$
$$= 2x-1+(1-x)^2[-2\rho x+\rho-2\rho+2\rho x+1+\rho] = x^2 > 0.$$

Similarly, $\partial N_P^+/\partial\lambda < 0$ follows from the sign of the determinant

$$\begin{vmatrix} -1 & 1 \\ 2\rho(1-x)^2-(1+\rho)(1-x) & 1-\rho(1-x)^2 \end{vmatrix}$$
$$= -1+\rho(1-x)^2-2\rho(1-x)^2+(1+\rho)(1-x) = -1+(1-x)(1+\rho x) = x(\rho-1-\rho x) < 0.$$

Turning now to the last result in the proposition, each active agent now has probability x_+^A (given by (18)) of being informed of e_+ when it occurs, in which case the high type will switch to $a=1$; therefore, $\bar{e} = (\rho/2)f_+e_+x_+^A$. The formula for x_+^A derived above shows that it is a rational fraction in λ, with determinant equal to $(1-x)$ times

$$\begin{vmatrix} -2 & 2-x \\ -(1+\rho)+2(1-x)\rho & 1-\rho(1-x)^2 \end{vmatrix}$$
$$= 2\rho(1-x)^2-2+2(1+\rho)-4(1-x)\rho-x(1+\rho)+2x(1-x)\rho$$
$$= 2\rho(1-2x)-2\rho+4\rho x+x(\rho-1) = x(\rho-1) < 0.$$

Therefore, x_+^A and $\bar{e}$ are both decreasing in λ. To show the corresponding results with respect to x, note first that $1/N_A^+$ is proportional to $1/(1-x)-(1+\rho)\lambda+(1-x)\rho(2\lambda-1)$, whose derivative has the sign of $1-(1-x)^2\rho(2\lambda-1) > 0$. Therefore N_A^+ is decreasing in x, and then a fortiori so is $N_P^+ = [(1-x)(1-\lambda)/[1-(1-x)\lambda]\,\rho(1+N_A^+)$. Turning finally to the variations of $\bar{e} = (\rho/2)f_+e_+x_+^A$, we compute

$$\begin{aligned} Z^2\frac{\partial x_+^A}{\partial x} &= [(2\lambda-1)(x-1)+1+(2\lambda-1)x]\left[1-(1-x)(1+\rho)\lambda+(1-x)^2\rho(2\lambda-1)\right] \\ &\quad - x[(2\lambda-1)(x-1)+1][\lambda(\rho+1)+\rho(2x-2)(2\lambda-1)] \\ &= [2(2\lambda-1)x)+2(1-\lambda)][1-(1-x)(1+\rho)\lambda+(1-x)^2\rho(2\lambda-1)] \\ &\quad - x[(2\lambda-1)(x-1)+1][\lambda(\rho+1)+\rho(2x-2)(2\lambda-1)]. \end{aligned}$$

The term in x^3 cancels out, leaving a polynomial $P(x) = Ax^2 + Bx + C$ with

$$\begin{aligned}
A &= (4\lambda - 2)\big[\lambda(\rho+1) - 2\rho(2\lambda-1)\big] \\
&\quad - (2\lambda-1)\big[\lambda(\rho+1) - 2\rho(2\lambda-1)\big] + \rho(2\lambda-1)(2\lambda-2) \\
&= \lambda(1-\rho)(2\lambda-1), \\
B &= (4\lambda-2)\big[\rho(2\lambda-1) - \lambda(\rho+1) + 1\big] \\
&= -2(1-\rho)(2\lambda^2 - 3\lambda + 1), \\
C &= -(2\lambda-2)\big[\rho(2\lambda-1) - \lambda(\rho+1) + 1\big] \\
&= 2(1-\rho)(1-\lambda)^2.
\end{aligned}$$

It is monotonic in x, since

$$\frac{P'(x)}{2(1-\rho)} = \lambda(2\lambda-1)\,x - \left(2\lambda^2 - 3\lambda + 1\right) = (2\lambda-1)[1 + \lambda(1-x)].$$

Moreover,

$$P(0) = C > 0, \qquad \frac{P(1)}{1-\rho} = \lambda(2\lambda-1) - 2\left(2\lambda^2 - 3\lambda + 1\right) + 2(1-\lambda)^2 = \lambda > 0,$$

therefore $P(x) > 0$ for all $x \in [0,1]$, implying the desired result. ■

Proof of Proposition 3 For negative signals, x_-^P and x_-^A are given by (17), independently of the whether the equilibrium is one with $a_H(\emptyset) = 1$ or $a_H(\emptyset) = 0$. It is immediate to see that x_-^P is decreasing in λ and x_-^A increasing and that their ratio $(2\lambda-1)x + 2(1-\lambda)$ takes values of $2 - x < 1/x$, 1 and x at $\lambda = 0$, $1/2$ and 1 respectively.

Turning now to positive signals, we first show that x_+^P is increasing in λ; indeed, the determinant equals $1 - x$ times

$$\begin{aligned}
&\begin{vmatrix} -2\rho & 1 + (1-x)\rho \\ 2\rho(1-x) - (1+\rho) & 1 - \rho(1-x)^2 \end{vmatrix} \\
&= -2\rho + \rho x + 1 + \rho^2 - \rho^2 x = (1-\rho)^2 + \rho x(1-\rho) > 0.
\end{aligned}$$

Next, from (A.1)-(A.2), we have:

$$\frac{x_+^P}{x_+^A} = \frac{1 - (1-x)\rho(2\lambda-1)}{1 - (1-x)(2\lambda-1)}, \tag{A.6}$$

which also increases in λ, and hence a fortiori so does x_+^P. Denoting $y \equiv 1 - x$, the ratio starts from $(1 + y\rho)/(1+y) < 1$ at the origin, reaches 1 at $\lambda = 1/2$ and continues rising to $(1 - y\rho)/(1-y) > 1$ at $\lambda = 1$. Noting that

$$\frac{1+y\rho}{1+y} \times \frac{1-y\rho}{1-y} = \frac{1 - y^2\rho^2}{1 - y^2} > 1$$

completes the proof. ■

Proof of Lemma 1 (1) Let $F_1 \preccurlyeq_{SOSD} F_2$ and $F_1(e^\dagger) \leq F_2(e^\dagger)$, and denote $\hat{F}_1$ and $\hat{F}_2$ the truncations of F_1 and F_2 respectively to $[0, e^\dagger]$. We have

$$E_{\hat{F}_2} - E_{\hat{F}_1} = \int_0^{e^\dagger} \left[\frac{F_1(z)}{F_1(e^\dagger)} - \frac{F_2(z)}{F_2(e^\dagger)}\right] dz \geq \frac{1}{F_1(e^\dagger)} \int_0^{e^\dagger} \left[\hat{F}_1(z) - \hat{F}_2(z)\right] dz \geq 0,$$

where the last inequality follows from $F_1 \preccurlyeq_{SOSD} F_2$. Thus $\mathcal{M}^-_{F_2}(e^\dagger) = E_{\hat{F}_2} \geq E_{\hat{F}_1} = \mathcal{M}^-_{F_1}(e^\dagger)$. For top-heaviness, now let $F_1(e^\dagger) \geq F_2(e^\dagger)$. We have pansion

$$\begin{aligned} \mathcal{M}^+_{F_1}(e^\dagger) - \mathcal{M}^+_{F_2}(e^\dagger) &= \int_{e^\dagger}^1 \left[\frac{1 - F_1(z)}{1 - F_1(e^\dagger)} - \frac{1 - F_2(z)}{1 - F_2(e^\dagger)}\right] dz \\ &\geq \frac{1}{1 - F_1(e^\dagger)} \int_{e^\dagger}^1 [F_2(z) - F_1(z)]\, dz \geq 0, \end{aligned}$$

where the last inequality follows from $F_1 \preccurlyeq_{SOSD} F_2$.

(2) Let X_1 and X_2 be random variables distributed on $[0, 1]$ with distribution functions F_1 and F_2 respectively. For any cutoff $e^\dagger \in [0, 1]$, integration by parts yields:

$$\mathcal{M}^-_{F_1}(e^\dagger) - e^\dagger = -\frac{\int_0^{e^\dagger} F_1(z)\, dz}{F_1(e^\dagger)} = -\left(\frac{\partial}{\partial e^\dagger}\left[\ln \int_0^{e^\dagger} F_1(z)\, dz\right]\right)^{-1}$$

Thus, $E[X|X \leq e^\dagger] \leq E[Y|Y \leq e^\dagger]$ if and only if the ratio $\int_0^{e^\dagger} F_1(z)dz / \int_0^{e^\dagger} F_2(z)dz$ (and therefore also its log) is strictly decreasing in $e^\dagger$. It is well-known that a sufficient condition is that F_2/F_1 be increasing in $e^\dagger$, for which it suffices in turn that f_2/f_1 have the same property. ■

Proof of Proposition 4 There are two cases to consider.

1. Reputation-enhancing excuses. Consider first the conditions for an equilibrium in which the high type searches more, $x_L \leq x_H$. Thus $\hat{v}_D \geq \bar{v}$ and $\hat{e} \geq e^*$ by (28), while (30)-(31) imply that $x_L \leq x_H$ if an only if:

$$\mathcal{M}^-(\hat{e}) v_H \leq c - \mu(v_H - v_L). \tag{A.7}$$

This condition will hold if $F(e)$ is sufficiently bottom-heavy, and fail if it is sufficiently top-heavy. Indeed, in the first case $\mathcal{M}^-(\hat{e})v_H$ decreases toward $0 \leq v_L < c - \mu(v_H - v_L)$, whereas in the latter it increases toward $v_H \hat{e} = c - \mu(v_H - \hat{v}_D) > c - \mu(v_H - v_L)$.

Although $\hat{e}$ itself varies with F, a *sufficient* condition that *precludes* any equilibrium with $x_H \geq x_L$, or equivalently $\hat{e} \geq e^*$, is $\mathcal{M}^-(e^*)v_H > c - \mu(v_H - v_L)$, which involves only exogenous

parameters. It will hold if F is insufficiently bottom-heavy, or too top-heavy.[27] Rewriting the inequality slightly using (3) yields the reverse of (33).

2. Reputation-tarnishing excuses. For an equilibrium in which it is the low type who searches more for excuses, $x_L \geq x_H$, hence $\hat{v}_D \leq \bar{v}$, $\hat{e} \leq e^*$ and (A.7) is reversed:

$$\mathcal{M}^-(\hat{e})v_H \geq c - \mu(v_H - v_L), \tag{A.8}$$

which will hold when F is sufficiently top-heavy ($\mathcal{M}^-(\hat{e})$ close to $\hat{e}$, meaning that F has relatively little mass below $\hat{e}$), or more generally not too bottom-heavy (which would make $\mathcal{M}^-(\hat{e})$ close to zero). In particular, a *sufficient* condition on exogenous parameters that *precludes* any such equilibrium is $\mathcal{M}^-(e^*)v_H < c - \mu(v_H - v_L)$, which holds when F is insufficiently top-heavy, or too bottom-heavy. Rewriting the inequality slightly using (3) yields (33).

It only remains to prove that an equilibrium with $a_H(\emptyset) = 1$ exists whenever (29) is satisfied. Equation (28) maps each $\hat{v}_D \in [v_L, v_H]$ into a unique cutoff $\hat{e} \in (0, 1]$, where $\hat{e} > 0$ follows from (2). To any such $\hat{e}$, equations (30)-(31) then associate a unique $(x_H, x_L) \in [0, 1]^2$, with $x_H > 0$ since $F(\hat{e}) > 0$ and $\mathcal{M}^-(\hat{e}) < \hat{e}$ due to f having full support. To any such pair, finally, (27) associates a new $\hat{v}'_D \in [v_L, v_H]$. Each of these mappings is continuous (the last one since $x_H > 0$), hence by Brouwer's theorem their composite has a fixed point ($\hat{v}_D = \hat{v}'_D$). ■

Proof of Proposition 6 Each type's expected utility from a search intensity x are now

$$\begin{aligned} U_H(x) &= -\psi(x) + x\left[F(\hat{e})\mu\hat{v}_D + \int_{\hat{e}}^1 (v_H e - c + \mu v_H) dF(e)\right] + (1-x)\int_0^1 \mu\hat{v}_{ND} dF(e) \\ &= -\psi(x) + x\mu(\hat{v}_D - \hat{v}_{ND}) \ + x\int_{\hat{e}}^1 v_H(e - \hat{e}) dF(e) + \mu\hat{v}_{ND}, \\ U_L(x) &= -\psi(x) + xF(\hat{e})\mu\hat{v}_D + [1 - xF(\hat{e})]\,\mu\hat{v}_{ND}, \end{aligned}$$

leading to the stated first-order conditions. It remains to prove that an equilibrium with $a_H(\emptyset) = 0$ exists when e_0 is low enough. Equation (28) again maps each $\hat{v}_D \in [v_L, v_H]$ into a unique cutoff $\hat{e} \in (0, 1]$. To any such $\hat{e}$, equations (36)-(37) now associate a unique pair $(x_H, x_L) \in [0, 1]^2$, with $x_H > x_L \geq 0$, as noted in the text. To any such pair, finally, (34) associates a new value $\hat{v}'_{ND} \in [v_L, \bar{v})$. Moreover, each of these mappings is continuous (the last one since $x_L < 1$), hence by Brouwer's theorem their composite has a fixed point $\hat{v}_{ND} = \hat{v}'_{ND}$ in $[v_L, \bar{v})$. For $v_H(e_0 - e^*) < \mu(v_L - \bar{v}) = -\mu\rho(v_H - v_L)$, moreover, equation (35) must then hold, so all equilibrium conditions are satisfied. ■

Proof of Proposition 7 Let $\underline{e}_0$ be the value of e_0 that makes (29) an equality; for all $e \geq e_0$, there exists an equilibrium with $a_H(\emptyset) = 1$. Turning to conditions for an equilibrium,

[27] This can be illustrated with specific distributions: (a) Let F have an atom of mass q at $e = 0$ and uniform density $1 - q$ on $[0, 1]$. Thus q directly measures bottom–heaviness, and $\mathcal{M}^-(e) = (e^*)^2/[2e^* + 2q/(1-q)]$. It is then easily seen that the sufficient condition becomes $q \leq q^*$, for some $q^* < 1$. Moreover, $q^* > 0$ if and only if $v_H e^* < 2[c/\beta - \mu(v_H - v_L)]$, or equivalently $\mu(1+\rho)(v_H - v_L) < c/\beta$. One could more generally take an atom at 0 or some $\underline{e} << e^*$ and the remaining mass distributed according to any continuous density over $[0, 1]$. (b) Consider now a top-heavy distribution, $f(e) = (1+\gamma)e^\gamma$, $\gamma \geq 0$, for which $\mathcal{M}^-(e) = e(1+\gamma)/(2+\gamma)$. The condition holds for $\gamma \geq \gamma^*$, where $\gamma^* < +\infty$. Moreover, $\gamma^* > 0$ under the same condition as $q^* > 0$ in the previous example. Case 2 below conversely corresponds to $q \geq q^*$ or $\gamma \leq \gamma^*$ in examples (a)-(b).

let $\hat{v}_{ND}(e_0) \in [v_L, \bar{v})$ denote any fixed point of the mapping defined by equations (28), (36)-(37) and (34); we saw in the proof of Proposition 6 that such a fixed point always exists, and that it defines an equilibrium if an only if $v_H(e_0 - e^*) \leq -\mu\left[\bar{v} - \hat{v}_{ND}(e_0)\right]$, which corresponds to condition (35). Let us now show that, as e_0 tends to $\underline{e}_0$ from above, $\hat{v}_{ND}(e_0)$ remains bounded away from v_L, which will imply that there exists a nonempty range $(\underline{e}_0, \bar{e}_0)$ in which $\mu(\bar{v} - v_L) < v_H(e_0 - e^*) < -\mu\left[\bar{v} - \hat{v}_{ND}(e_0)\right]$, so that both equilibria coexist. From (34), it suffices that $x_H(e_0)$ remain bounded away from 1, and from (36) this is ensured as long as $\psi'(1) = +\infty$, since the right-hand side of (36) is bounded above by $\mu(v_H - v_L) + v_H[\mathcal{M}^+(\hat{e}) - \hat{e}] < \mu(v_H - v_L) + v_H$. ■

Supplementary Online Appendix

Refinements and Uniqueness under Pure Reputation Concerns in Section 4.

Denote by (x_H, x_L) be the probabilities (exogenous or endogenous) with which each type obtains some narrative e drawn from $[0,1]$ according to F, $a_H(e)$ the action choice of the informed high type, and denote $A_1 \equiv \{e|a_H(e)=1\}$ and $A_0 \equiv \{e|a_H(e)=0\}$. For values of $e \in A_0$, let D_i denote the subset disclosed in equilibrium by type $i = H, L$, and N those disclosed by neither. For any subset $X \subset [0,1]$, let $P(X)$ be the probability measure of X according to the distribution $F(e)$. We first establish a series of claims pertaining to any Perfect Bayesian Nash equilibrium in which off-equilibrium beliefs are restricted only by the elimination of strictly dominated strategies.

Claim 1. $D_L = D_H \equiv D \subseteq A_0$.

Proof. For the high type choosing $a = 1$ is perfectly revealing, so disclosure has no benefit and involves a small cost, and is thus a strictly dominated strategy. For any $e \in A_1$, disclosure would then be interpreted as coming from the low type for sure, resulting in reputation v_L and involving a cost, which is dominated by nondisclosure. Therefore, $D_H \subseteq A_0$.

Next, if some e were disclosed only by the low type it would yield minimal reputation v_L and involve a cost, so it must be that $D_L \subset D_H$. If some e was disclosed only by the high type it would yield maximal reputation v_H, so the low type would imitate, unless $\hat{v}_{ND}$ was equal to v_H; that, however, would require that the low type always disclose, a contradiction.

Claim 2. *For any $e \in D$, beliefs following $a = 0$ and disclosure are independent of e, which we denote as $\hat{v}(e) \equiv \hat{v}_D$, and given by the likelihood ratio:*

$$\hat{L}_D = \frac{\rho}{1-\rho}\frac{x_H}{x_L}. \tag{OA.1}$$

As to beliefs $\hat{v}_{ND}$ following $a = 0$ and no disclosure, they are given by

$$\hat{L}_{ND} = \frac{\rho}{1-\rho}\frac{1 - x_H + x_H P(N)}{1 - x_L + x_L\left[P(N) + P(A_1)\right]}. \tag{OA.2}$$

Furthermore, the following three properties are equivalent:

(i) $\hat{v}_D < \hat{v}_{ND}$

(ii) $x_H - x_L + x_H x_L P(A_1) > 0$

(iii) $\hat{v}_{ND}$ *is increasing in* $P(N)$.

Proof: The constancy of $\hat{L}$ and $\hat{v}$ over all $e \in D$ follows from Claim 1 and the formulas for $\hat{L}_D$ and $\hat{L}_{ND}$ from Bayes' rule. Note, that for $e \notin D$, in contrast, any beliefs $\tilde{v}(e) \leq v_{ND}$ are generally allowed. Next, define the function

$$Q(Z) \equiv \frac{1 - x_H + x_H Z}{1 - x_L + x_L\left[Z + P(A_1)\right]},$$

and observe from (OA.2) that $\hat{L}_{ND} = Q(P(N))$. It is easily verified that Q is increasing in Z if condition (ii) holds, and decreasing when it is reversed. Note also, from (OA.1), that

$\hat{L}_D = Q(+\infty)$, which concludes the proof. [28]‖

Remark. The fact that $\hat{v}_{ND}$ is increasing in $P(N)$ whenever $\hat{v}_D > \hat{v}_{ND}$ is important is what precludes ruling out partial-disclosure equilibria ($D \subsetneq A_0$) by Pareto dominance. If both types were to coordinate on disclosure for any subset of N they would be better off for such realizations of e (reputation $\hat{v}_D > \hat{v}_{ND}$ rather than $\tilde{v}(e) \leq v_{ND}$) but worse off under all cases of non-disclosure (a lower $\hat{v}_{ND}$), and in particular in the "unavoidable cases" where no narrative is received or found. With disclosure of some values of e his precluded by very unfavorable out-of-equilbrium beliefs, moreover, the high type may prefer to choose $a = 1$ even at relatively low values of e, meaning that his equilibrium choice of a is no longer a threshold rule.

Refinement assumption. Suppose that $e \in N$, and deviation is nonetheless observed. Given that they care equally about reputation, neither type gains or loses more than the other from any given off-path belief $\tilde{v}(e)$. There is thus no reason for observers to infer that the deviation was more likely to come from the low type, *controlling* for each-type's likelihood of being informed in the first place. Yet, as we show below, that is precisely what is needed to sustain equilibria with nonempty N. Conversely, the natural restriction that disclosure leads to the same belief $\hat{v}_D$ (reflecting the probabilities of each type being informed) off and on the equilibrium path rules out all but the threshold-type equilibrium we have focussed on in the main text.

Claim 3. *(i) Let x_H and x_L be endogenously chosen, at cost $\psi(x)$. In any equilibrium, it must be that $\hat{v}_D > \hat{v}_{ND}$; the other conditions in Claim 2 must therefore hold as well, and some disclosure must occur in equilibrium: $D \neq \emptyset$. (ii) These same properties hold when (x_L, x_L) are exogenous, provided $x_H \geq x_L$ and $x_H > 0$.*

Proof: (i) If $\hat{v}_D \leq \hat{v}_{ND}$, type L never discloses (whether $e \in D$ or not), as the resulting reputation is bounded by $\hat{v}_{ND}$ and there is a slight cost of disclosure. It must then be that $x_L = 0$, as acquiring costly but useless information would be a a strictly dominated strategy. If $x_H > 0 = x_L$ then disclosure reveals the H type, $\hat{v}_D = v_H > \hat{v}_{ND}$, hence a contradiction. If $x_H = 0 = x_L$ then $v_{ND} = \bar{v}$; information has no reputation value but retains a strictly positive decision value for the H type: since both $e < e^*$ and $e > e^*$ have positive probability (as F has full support), he is willing to pay a positive cost just to set a_H optimally (without disclosing). Therefore $x_H > 0$, a contradiction. (ii) The properties follow directly from Claim 2(ii). ‖

Proposition 8. *Assume that, following $a = 0$ and the unexpected disclosure of some $e \in N$, out-of equilibrium beliefs are the same $\hat{v}_D$ as would follow $a = 0$ and any $e' \in D$. In equilibrium, $A_1 = (\hat{e}, 1]$, $A_0 = [0, \hat{e}]$ and $D \in \{\emptyset, A_0\}$, with the cutoff $\hat{e}$ given by:*

$$v_H \hat{e} - c + \mu(v_H - \max\{\hat{v}_D, \hat{v}_{ND}\}) \equiv 0.$$

Under either condition in Claim 3 $\hat{v}_D > \hat{v}_{ND}$, so this reduces to (28), and $D = A_0, N = \emptyset$.

Proof. If an informed agent chooses $a = 0$ and discloses he gets reputation $\hat{v}_D$, independently of the disclosed e, and whether $e \in D$ or $e \in N$. The results follow immediately. ■

[28] The fact that $\hat{v}_{ND}$ is increasing in $P(N)$ whenever $\hat{v}_D > \hat{v}_{ND}$ is what precludes ruling out partial-disclosure equilibria ($D \subsetneq A_0$) by Pareto dominance. If both types were to coordinate on disclosure for any subset of N, they would be better off for such realizations of e (reputation $\hat{v}_D > \hat{v}_{ND}$ rather than $\tilde{v}(e) \leq v_{ND}$) but worse off under all cases of non-disclosure (a lower $\hat{v}_{ND}$), and in particular whenever no narrative is found. With disclosure of some values of e thus precluded by unfavorable off-path beliefs, moreover, the high type may prefer to choose $a = 1$ even at relatively low values of e, meaning that his equilibrium choice of a is no longer a threshold rule.

References

Aina, C. (2023) "Tailored Stories." Working paper, Barcelona School of Economics. https://chiaraaina.github.io/research/tailored_stories/

Aina, C. and F. H. Schneider (2025) "Weighting Competing Models." CESifo Working Paper No. 11757; CEBI Working Paper 25-04. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5200330

Akerlof, R. and R. Shiller (2015) *Phishing for Phools*. Princeton, NJ: Princeton University Press.

Ambrus, A., Azevedo, E. and Y. Kamada (2013) "Hierarchical Cheap Talk," *Theoretical Economics*, 8: 233–261.

Andre, P., I. Haaland, C. Roth, M. Wiederholt and J. Wohlfart (forthcoming) "Narratives about the Macroeconomy," *Review of Economic Studies*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4947636

Andre, P., C. Pizzinelli, C. Roth and J. Wohlfart (2022) "Subjective Models of the Macroeconomy: Evidence from Experts and Representative Samples," *Review of Economic Studies*, 89(6): 2958–2991.

Barrera, O., Guriev, S., Henry, E. and E. Zhuravskaya (2020) "Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics," *Journal of Public Economics*, 182, 104123.

Barron, K. and T. Fries (2023) "Narrative Persuasion." CESifo Working Paper No. 10206; WZB Discussion Paper SP II 2023-301r. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4329465

Barron, K., H. Harmgart, S. Huck, S. O. Schneider and M. Sutter (2023) "Discrimination, Narratives, and Family History: An Experiment with Jordanian Host and Syrian Refugee Children," *Review of Economics and Statistics*, 105(4): 1008–1016.

Bartling, B. and U. Fischbacher (2012) "Shifting the Blame: On Delegation and Responsibility," *Review of Economic Studies*, 79(1): 67–87.

Bartling, B., V. Valero, R. A. Weber and L. Yao (2024) "Public Discourse and Socially Responsible Market Behavior," *American Economic Review*, 114(10): 3041–3074.

Bénabou, R. and J. Tirole (2006) "Incentives and Prosocial Behavior," *American Economic Review*, 96(5): 1652–1678.

Bénabou, R. and J. Tirole (2011a) "Identity, Morals and Taboos: Beliefs as Assets," *Quarterly Journal of Economics*, 126(2): 805–855.

Bénabou, R. and J. Tirole (2011b) "Laws and Norms." NBER Working Paper 17579, November.

Bénabou, R. and J. Tirole (2016) "Mindful Economics: The Production, Consumption and Value of Beliefs." *Journal of Economic Perspectives*, 30(3): 141–164.

Bénabou, R., A. Falk and J. Tirole (2019) "Narratives, Imperatives, and Moral Reasoning." NBER Working Paper No. 24798.

Besley, T. and A. Brzezinski (2025) "The Political Economy of Neoliberal Narratives." CEPR Discussion Paper No. 20410. https://cepr.org/publications/dp20410

Bloch, F., R. Kranton and G. Demange (2018) "Rumors and Social Networks," *International Economic Review*, 59(2): 421–448.

Bruner, J. (1991) "The Narrative Construction of Reality," *Critical Inquiry*, 18(1): 1–21.

Bursztyn, L., G. Egorov and S. Fiorin (2019) "From Extreme to Mainstream: The Erosion of Social Norms." University of Chicago mimeo.

Campbell, A., D. J. Thornton and Y. Zenou (2025) "Strategic Influence: The Diffusion of Prosocial and Antisocial Behaviors." CEPR Discussion Paper No. 20425. https://cepr.org/publications/dp20425

Charles, C. and C. Kendall (2022, revised 2025) "Causal Narratives." NBER Working Paper No. 30346. https://www.nber.org/papers/w30346

Dal Bó, E. and P. Dal Bó (2014) "'Do the Right Thing': The Effects of Moral Suasion on Cooperation," *Journal of Public Economics*, 117: 28–38.

Darley, J. M. and B. Latané (1968) "Bystander Intervention in Emergencies: Diffusion of Responsibility," *Journal of Personality and Social Psychology*, 8(4): 377–383.

DellaVigna, S., List, J. and U. Malmendier (2012) "Testing for Altruism and Social Pressure in Charitable Giving," *Quarterly Journal of Economics*, 127: 1–56.

Dewatripont, M. and J. Tirole (2024) "The Morality of Markets and Organizations," *Journal of Political Economy*, 132(8): 2655–2694.

Ditto, P. H., Pizarro, D. A. and D. Tannenbaum (2009) "Motivated Moral Reasoning," in D. M. Bartels, C. W. Bauman, L. J. Skitka and D. L. Medin (eds.), *The Psychology of Learning and Motivation*, vol. 50. Burlington, VT: Academic Press, pp. 307–338.

Eliaz, K. and R. Spiegler (2020) "A Model of Competing Narratives," *American Economic Review*, 110(12): 3786–3816.

Eliaz, K. and R. Spiegler (2024) "News Media as Suppliers of Narratives (and Information)." Working paper, arXiv:2403.09155. https://arxiv.org/abs/2403.09155

Ellingsen, T. and M. Johannesson (2008) "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98(3): 990–1008.

Exley, C. L. (2016) "Excusing Selfishness in Charitable Giving: The Role of Risk," *Review of Economic Studies*, 83(2): 587–628.

Falk, A. and N. Szech (2013) "Morals and Markets," *Science*, 340: 707–711.

Falk, A., T. Neuber and N. Szech (2020) "Diffusion of Being Pivotal and Immoral Outcomes," *Review of Economic Studies* (forthcoming).

Foerster, M. and J. J. van der Weele (2021) "Casting Doubt: Image Concerns and the Communication of Social Impact," *Economic Journal*, 131(639): 2887–2919.

Galeotti, A., C. Ghiglino and F. Squintani (2013) "Strategic Information Transmission in Networks," *Journal of Economic Theory*, 148(5): 1751–1769.

Graeber, T., C. Roth and F. Zimmermann (2024) "Stories, Statistics, and Memory," *Quarterly Journal of Economics*, 139(4): 2181–2225.

Hagenbach, J. and F. Koessler (2010) "Strategic Communication Networks," *Review of Economic Studies*, 77(3): 1072–1099.

Haidt, J., J. Graham and C. Joseph (2009) "Above and Below Left-Right: Ideological Narratives and Moral Foundations," *Psychological Inquiry*, 20(2–3): 110–119.

Hillenbrand, A. and E. Verrina (2022) "The Asymmetric Effect of Narratives on Prosocial Behavior," *Games and Economic Behavior*, 135: 241–270.

Juille, T. and D. Jullien (2017) "Narrativity and Identity in the Representation of the Economic Agent," *Journal of Economic Methodology*, 24(3): 274–296.

Michalopoulos, S. and M. M. Xue (2021) "Folklore," *Quarterly Journal of Economics*, 136(4): 1993–2046.

Mukand, S. and D. Rodrik (2018) "The Political Economy of Ideas: On Ideas Versus Interests in Policymaking." NBER Working Paper No. 24467.

Roos, M. and M. Reccius (2024) "Narratives in Economics," *Journal of Economic Surveys*, 38(2): 303–341.

Schwartzstein, J. and A. Sunderam (2021) "Using Models to Persuade," *American Economic Review*, 111(1): 276–323.

Schwartzstein, J. and A. Sunderam (2024) "Sharing Models to Interpret Data," *Harvard Business School Working Paper*, 25-011.

Shiller, R. J. (2017) "Narrative Economics," *American Economic Review*, 107(4): 967–1004.

Somers, M. and F. Block (2005) "From Poverty to Perversity: Ideas, Markets, and Institutions over 200 Years of Welfare Debate," *American Sociological Review*, 70: 260–287.

Spiegler, R. (2016) "Bayesian Networks and Boundedly Rational Expectations," *Quarterly Journal of Economics*, 131(3): 1243–1290.

Sykes, G. M. and D. Matza (1957) "Techniques of Neutralization: A Theory of Delinquency," *American Sociological Review*, 22(6): 664–670.

Tirole, J. (2026) "Safe Spaces: Shelters or Tribes." forthcoming, Journal of Political Economy.