Narratives, Imperatives, and Moral Reasoning

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

This version: December 31, 2018[4]

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

[2] Institute on Behavior and Inequality (briq) and Department of Economics, University of Bonn.

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

[4] We are thankful for valuable comments and suggestions from Johannes Abeler, Alberto Alesina, Nageeb Ali, Daniel Chen, Thomas Graeber, Johannes Hermle, Ian Jewitt, Alessandro Lizzeri, Stephen Morris, Christopher Roth, Gilles Saint-Paul, Paul Seabright, Nora Szech, 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, responsibilizing narratives raise both direct and reputational stakes, fostering prosocial behavior. Such rationales diffuse along a random network, through costly signaling and strategic disclosure. Norms of conduct and discourse, average compliance, and belief polarization reflect local correlation in types' tradeoffs between reputation and influence concerns. Imperatives (general precepts) constitute alternative modes of moral influence. We analyze the sources of their legitimacy, then their costs and benefits relative to narratives.

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 reasoning

What is the moral thing to do? This paper does not (of course) aim to answer that question, but instead to analyze the production and circulation of *arguments* seeking to justify different courses of action on the basis of morality. Such appeals to notions of "right or wrong" are central to public goods provision and the upholding of norms. They also pervade social and political discourse, often outweighing any arguments of economic efficiency: bans on "immoral" transactions or markets, trade policy, undeservedness of some group, or populism.

Some moral arguments provide reasons for what one "should do," or conversely for acting according to self-interest, under specific circumstances. Others are instead "fiat" prescriptions, dictating a fixed behavior across most situations, without explaining why. We refer to them as moral *narratives* and *imperatives,* respectively. They operate through very different mechanisms but serve similar ends, and moral discourse is even epitomized by its back and forth between consequentialist and deontological reasoning. It is therefore important to analyze both in an integrated framework, while at the same time identifying their distinctive features and properties. For narratives, we emphasize in particular the issue of *viral transmission:* what types of social structures lead exculpatory versus responsibilizing rationales to spread widely, or remain clustered within subgroups? For imperatives, the central question is what makes them work: who has the *moral legitimacy* to issue edicts that others will obey, and when will this be more effective than communicating specific reasons?

1.2 Narratives and imperatives: an economic view

Narratives are stories people tell themselves, and each other, to make sense of human experience– that is, to organize, explain, justify, predict and influence its course. They are "instrument[s] of mind in the construction of reality" (Bruner 1991), and viewed as central features of all societies by many disciplines including anthropology, psychology, sociology, history and the humanities.[1]

Given such a broad concept, it is useful to first distinguish two main types, and roles, of narratives. A first one, which we shall not directly address, is that of *narratives as sense-making:* people constantly seek to "give meaning" to disparate, sometimes even random events (Karlsson et al. 2004, Chater and Loewenstein 2010). This drive reflects a need for predictability, serving both planning and anxiety-reduction purposes, and is linked to evolutionary benefits of pattern seeking. The second sense, which we emphasize here, is that of *narratives as rationales or justifications.* These may be objectively relevant facts, but also political and social slogans ("Never again"), advertising pitches ("Because you are worth it"), or rationalizations such as "They are not making any more land" for real-estate bubbles (Shiller 2017). The most important narratives, however, pertain to actions with *moral implications,* namely those involving externalities and reputational concerns. It is on such rationales for what one "ought to do" (or not) that we focus. Accordingly, we define a moral narrative as any signal, story, or heuristic that can potentially alter an agent's beliefs about the tradeoff between private benefits and

[1] In McAdams' (1985, 2006) psychological framework, for instance, personality consists of three tiers: dispositional traits, contextual adaptations such as beliefs or values, and "life stories" providing an overall sense of meaning, unity and purpose.

social costs (or the reverse) faced by a decision-maker. The latter could be the agent himself, someone he observes, or someone he seeks to influence. The rationale may be received exogenously, searched for and thought of by the individual himself, or strategically communicated by another person.

Having the potential to alter beliefs does not necessarily require a story to be true or relevant, nor that receivers respond to it with fully rational (Bayesian) updating. Such can of course be the case, as with hard evidence on (say) pollution or corruption, and our framework admits an entirely rational reading in which only "true" signals count. All that matters in practice, however, is that there be a *perceived* "grain of truth" prompting persuasion. Some of the most influential narratives, conveying negative ethnic and gender stereotypes, are simply wrong. Even a real fact can provide a very incomplete picture, used to support a misleading conclusion: "This year's frigid winter proves that global warming is a hoax." Vivid life experiences, simple, striking arguments and emotion-laden cues are especially likely to be overweighted relative to "cold" statistical facts and to facilitate viral, word-of-mouth transmission. Cognitive biases and motivated reasoning also offer many avenues for narratives to "work" where, under full rationality, they should not.

Whereas good narratives either are, or *de facto* act like hard information, *imperatives* are entirely soft messages of the type "thou shalt (not) do this," seeking to constrain behavior without offering any reason other than a tautological "that is just wrong," or "because I say so". Thus, while a narrative can by itself influence beliefs and actions, independently of where it came from (rumor, story or picture "going viral" on social media), imperatives are fundamentally relationship-dependent: whether such mandates are obeyed or ineffective depends on the extent to which their author is regarded as trustworthy and benevolent. A second key distinction is that, whereas narratives often involve fine situational distinctions ("casuistry"), imperatives allow no or very little adjustment for contingencies. This stark representation of the two forms of moral discourse is of course a simplifying first step. In practice, most arguments lie along a continuum between these polar opposites, or explicitly combine them.

1.3 Formalization and main results

Following the utilitarian tradition, in which morality is typically described in terms of avoiding and preventing harm to others (Bentham 1789; Mill, 1861; Gert and Gert 2016), we define an action as moral if it produces a positive externality. Agents differ in their concern for others, may (but need not) have imperfect self-control, and derive reputational benefits from being perceived, or seeing themselves, as highly moral types –in line with ample evidence that people strive to maintain a positive self-concept and social image (e.g., Aquino and Reed 2002; Mazar et al. 2008; Monin and Jordan 2009). An agent is then more likely to act in the social interest the higher the perceived externality, his image concern and his willpower, and the lower his initial reputation. We discuss how these simple predictions match a wide range of experimental evidence from both psychology and economics.

This basic building block is then combined with *narratives*, introduced as disclosable rationales about the social implications of a person's actions. Abstracting from any specific channel, "rational" or "behavioral," through which different arguments may sway beliefs, we focus on why and how people use (invoke or withhold) them and what social norms emerge when both words and deeds convey information, as they do in practice. Two main types of arguments are thus relevant: by downplaying externalities or emphasizing personal costs, *negative narratives* or *excuses* allow an individual to behave selfishly while maintaining a positive self- and/or social image. Conversely, *positive narratives* or *responsibilities* increase the pressure to "do the right thing." We discuss a range of historical and experimental examples of both types: common "neutralization" rationales include denials of responsibility or injury and the derogation of victims, while classical "responsibilization" arguments involve appeals to empathy and imagined counterfactuals ("how would you feel in their place?", "what if everyone did this?").

Virality and network structure. The central focus of our analysis of narratives is their social-contagion aspect. To that end, we embed the basic framework into a linear network that stochastically mixes agents with different signaling and disclosure incentives. Each individual may observe what their predecessor did, receive a narrative from them, and transmit it to their successor; if he is among the "active" agents, he also chooses an action. Our running example is that of men and women in the workplace, how the former treat the latter, and the different arguments circulating about those behaviors. Other applications include majority and minority, or rich and poor. We obtain three sets of results.

First, the spread of opposing rationales through a population is driven by type-specific tradeoffs between what we term the *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. Conversely, sharing a responsibilizing narrative forces one 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 *complements,* and characterize what individuals relay or withhold depending on their moral types and reputational stakes.

Second, we determine how far each type of argument travels as a function of the network structure, and what type of discourse prevails. The social norm can be one in which either prosociality or selfishness is the default –what a moral type does when uninformed. In the first case, doing the right thing "goes without saying," whereas abstaining requires an excuse, so only negative narratives are used (when available). In the second, someone pursuing self-interest can plead ignorance, but having a justification is better, so those again circulate. But now so do positive narratives, propagated by high-morality actors (and by non-actors) seeking to induce others to behave responsibly; conversely, intentional "silence is complicity".

Third, we show that in either type of equilibrium, more *mixed interactions* between agents with differing reputational concerns *raise prosocial behavior.* People 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 responsibilizing ones. In the latter case so do high-morality actors, with complementary feedbacks. Turning from average behavior to dispersion, we also show that more random mixing within the network results in beliefs that are both less clustered and *less polarized.* Conversely, a high or even just predictable correlation pattern causes very different kinds of narratives to circulate in the two groups. In the lead example, men will share *more excuses* and rationalizations for behaviors that women will simultaneously view as *more inexcusable,* compared to what would occur in a more integrated setting.

Imperatives in a utilitarian framework. The other main form of moral argument is imperatives, and the key questions here are what makes people obey them and how their effectiveness compares to that of narratives. This leads us to expand the model's "influence" channel, but focusing now on communication in a single principal-agent dyad. The principal cares for the welfare of society at large and/or that of the agent, and she can issue either a narrative or an imperative. As before, narratives are signals, or messages interpreted as such, about the social or long-run impact of the agent's choices. In contrast, an imperative is a precept to act in a certain way, without looking to consequences for reasons.

The analysis reveals several important tradeoffs. On the cost side, mandates are effective only if issued by principals with *moral authority* (perceived competence and congruence), whereas rationales can persuade irrespective of their source. Imperatives are thus rarely used in the political arena –where narratives instead prevail– but common in parent-child interactions and religious writings, such as the Ten Commandments. Another restrictive feature of precepts and rules is that they impose rigidity on decision-making, leaving little room for adapting to contingencies. This is often identified as an important weakness of deontological reasoning, as in the case of Kant's imperative never to lie, even to a murderer at the door to save the life of a friend.[2] On the benefit side, imperatives are less vulnerable than justifications to the risk of misinterpretation, or counter-arguments. When effective, they also expand the range of externalities over which the principal can induce desired behaviors by the agent. This helps explain why imperatives typically consist of unconditional prescriptions with a large scope, such as not to lie, steal or kill. Relatedly, we show that the use of imperatives rather than rationales is more likely (up to a point) for agents who suffer from self-control problems.

1.4 Further results

Having emphasized the social transmission of narratives, we turn in the Supplementary Appendix to the "production side", by allowing an individual to engage in his own *search for reasons to act,* or not. Is having a rationale for self-interest then more indicative of a low-morality type who only looks for alibis, or a high-morality one who seeks to ascertain his responsibilities?

This depends, intuitively, on a comparison between: (i) the option values of discovering that the externality is substantially away from the prior mean, in either direction, versus: (ii) that of learning that it is simply low enough to provide an acceptable excuse. That tradeoff itself hinges on expected search intensities so, as we discuss in Section 3.5, endogenous *moral standards* emerge, characterizing how tolerant a group or society is of excuses: how strong they have to be, and how much stigma is borne by those who fail to produce a receivable one.

1.5 Related literature

The paper contributes to several lines of work. The first is the literature linking prosocial behavior with signaling or moral-identity concerns (e.g., Bénabou and Tirole 2006a, 2011a,b, Ellingsen and Johannesson 2008, Ariely et al. 2009, DellaVigna et al. 2012, Exley 2016, Grossman and van der Weele 2017). To the usual choice dimension of agents taking some costly

[2] "To be *truthful* (honest) in all declarations is therefore a sacred command of reason prescribing unconditionally, one not to be restricted by any conveniences." (Kant 1797, 8: 427). In the Appendix, we extend the model to explain why agents may *refrain from questioning* an imperative, even though this might yield good reasons why it is maladapted to current circumstances; see the end of Section 4 for a discussion.

action, we add the direct sharing of arguments and justifications. Thus, it is both *what people do* and *what they say* (or not) that determines how the audience judges them and responds.

This communication aspect relates the paper to public goods and learning in networks. Most of this literature features agents who learn mechanically from their neighbors and spontaneously emit information toward them (e.g., DeGroot model). Our paper is more connected to a recent strand in which communication is instead strategic (e.g., Hagenbach and Koessler 2010, Galeotti et al. 2013, Ambrus et al. 2013, Acemoglu and Jackson 2015, Bloch et al. 2016), and the first such one combining costly signaling with selective disclosure.

Because the arguments that individuals produce and circulate center here on social responsibility, the paper also clearly belongs to the fast-growing line of work exploring the interactions between morality, prices, and markets (e.g., Brekke et al. 2003, Roemer 2010, Falk and Szech 2013, 2017, Ambuehl 2016, Elias et al. 2016). The formal framework it develops, on the other hand, is much more generally applicable.

This work also ties into the literature on the cultural transmission of values, beliefs and norms (e.g., Bisin and Verdier 2001, Bénabou and Tirole 2006b, Tabellini 2008, Dohmen et al. 2012). Finally, the importance of informal "stories" in shaping economic actors' beliefs was emphasized by Akerlof and Shiller (2015) and Shiller (2017), and other papers are now also exploring the roles played by narratives, memes or myths (Juille and Jullien 2016, Mukand and Rodrik 2016, Barrera et al. 2017, Michalopoulos and Xue 2018, Eliaz and Spiegler 2018).

Most closely related is independent work by Foerster and van der Weele (2018a). Their model has two agents, each endowed with a prosociality type and an imperfect signal on the externality; they exchange cheap-talk messages about it, then both act. Image concerns create an incentive to understate externalities from behaving selfishly ("denial"), while the desire to make the other player behave better pushes towards exaggerating them ("alarmism"). This broadly parallels, in the form of distorted reports, how reputation and influence motives lead here to selective disclosure of negative versus positive narratives. The two models have different communication technologies, and our agents can also endogenously search for arguments. Most important, however, is the difference in questions investigated. Foerster and van der Weele show that any informative equilibrium must involve some denial, and that the latter may improve both parties' welfare. Our focus is on the dynamics of (serial) communication between many agents: strategic complementarity or substitutability, viral diffusion, group polarization, and the effects of social mixing. We also investigate imperatives, which are unidirectional cheap talk, and their tradeoffs with persuading through narratives.

The paper proceeds as follows. Section 2 introduces a simple setup in which moral values, esteem concerns and narratives jointly shape individual behavior. Section 3 embeds it into a random network to study how the diffusion of arguments, resulting norm, and belief polarization reflect the interplay of reputation and influence motives, filtered through social mixing. Section 4 turns to imperatives, first studying the sources of their legitimacy, then their costs and benefits, relative to moral narratives. Section 5 concludes. All proofs are in appendices.

2 Basic Model

2.1 Moral decisions and moral types

1. Preferences. The model's first element builds on Bénabou and Tirole (2006, 2011a). 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. Moral decisions may also involve internalities, due to weakness of will at the moment of choice; in such cases $c = c_0/\beta$, where c_0 is the ex-ante cost of "doing the right thing" and $\beta \leq 1$ is the individual's degree of self-control.[3]

Agents differ by their intrinsic motivation (or "core values") to act morally: 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 $\bar{v} = \rho v_H + (1 - \rho) v_L$. Note that these preferences are explicitly consequentialist: an agent's desire to behave prosocially is proportional to the externality he perceives his actions to have.

In addition to intrinsic fulfillment, 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. His utility level could also include direct benefits (if any) received from others' decisions, but he takes those as given.

It may seem that we only consider here behaviors that both actor and audience judge prosocial to some extent, and which the latter accordingly rewards with esteem: helping, not stealing from or exploiting others, etc. What of actions that are judged as moral by one group and immoral by another, because they perceive opposite externalities from them due to differing preferences or/and priors? Such polarized views (on abortion, guns, religion, politics, 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, yet we show in the Appendix how many cases still reduce to an "on net" unidimensional model very similar to the present one, which thus covers a much wider range of applications than initially appears.

2. Individual behavior. 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] The model also allows for $\beta > 1$, namely impulses to act selflessly rather than selfishly. Our reading of the evidence is that it goes mostly in the other direction, with most cases of "excessive" generosity resulting instead from strong self or social image concerns; see Section 2.2. The assumption that $e \leq 1$ could easily be relaxed.

Assumption 1.

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

The first inequality says that $a = 0$ is a strictly dominant strategy for the "immoral" v_L 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 v_H 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 easily leads to the following result:

Proposition 1 (determinants of moral behavior). *The moral type contributes if and only if $e > e^*$, where e^* is uniquely defined by*

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

Immoral behavior is encouraged by a low perceived social benefit e, a high personal cost c or low degree of self control β, and a weak reputational concern μ.[5]

We next discuss how these predictions align with a broad range of empirical evidence. The purpose is not to test this basic component of the model but to: (i) verify that it is empirically sound before proceeding to build further upon it; (ii) show how it can usefully organize a large number of seemingly disparate experimental findings. Readers primarily interested in the modeling of narratives and imperatives can proceed directly to Section 2.3.

2.2 Related evidence

1. Externality (e). That choices are generally sensitive to the implied social consequences is well documented in the literature on cooperation and voluntary contribution to public goods (e.g., Kagel and Roth 1995). In a study that cleanly disentangles higher external return (gain

[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 better for both types, and thus selected by our Pareto criterion.

for others) from internal cost (to the subject), Goeree et al. (2002) show that the two have opposite effects on the level of contributions. Likewise, charitable giving decreases when the risk of having no impact rises (Brock et al. 2013). In a field study, Gneezy et al. (2014) show that donations to charity decrease when overhead increases, and conversely they rise when potential donors are informed that those costs are already covered. Taking into account the magnitude of externalities is also central to the idea of "effective altruism," which calls for choosing those charitable donations with the highest social rate of return. We model agents' preferences in line with this notion and the above evidence, but also take note of two important types of insensitivity to consequences. One stems from impure altruism or "warm glow," where utility is derived from the act as such, not what it achieves (e.g., Andreoni 1989, 1990). The other, on which our companion paper (Bénabou et al. 2018) focuses, is a stated unwillingness to enter moral tradeoffs altogether, often referred to as deontological or Kantian reasoning.

2. Costs (c). That prosocial behavior responds to the personal cost involved is intuitive and would be the implication of most models, except when multidimensional signaling gives rise to a sufficiently strong crowding-out effect (downward-sloping supply), as in Bénabou and Tirole (2006a). In public goods games, for instance, the cost of providing a positive externality reduces the level of cooperation (Goeree et al. 2002, Gächter and Herrmann 2009), and the willingness to exert altruistic punishment decreases in the cost of sanctioning (Egas and Riedl 2008, Nikiforakis and Normann 2008). In Falk and Szech (2013), subjects could either kill a (surplus) mouse in return for money, or decline to. As the price offered rises so does the fraction willing to do the deed, although there remains some subset who refuse even at the maximum price, exhibiting the type of "deontological" behavior mentioned above.

3. Self-control ($\beta = c_0/c$). Morally demanding decisions often imply a tradeoff between immediate gratification and future consequences: guilt or pride, social reputation, or outright punishment. In particular, Gottfredson and Hirschi's (1990) influential "self-control theory of crime" appears well supported by empirical studies (e.g., Lagrange and Silverman 2006). In experiments, Martinsson et al. (2012) find that dictator-game participants who report generally having low self control make more selfish allocations, and Achtziger et al. (2015) find similar behavior when subjects are experimentally "ego depleted". Related experiments show that depleted self-control also fosters dishonesty (Gino et al. 2011, Mead et al. 2009) and undermines cooperation (Osgood and Muraven 2015). Neuroscientific evidence further suggests that an inhibition of self-control areas (dorsolateral prefrontal cortex) through transcranial magnetic stimulation induces more selfish behavior (Knoch et al. 2006).

4. Social and self-image concerns (μ). Increased visibility is predicted to induce more moral behaviors, as indeed found in many contexts, ranging from charitable contributions (e.g., Ariely et al. 2009, Ashraf et. al. 2012) to public goods provision (Algan et al. 2013), voting (Gerber et al. 2008) and blood donations (Lacetera and Macis 2012). The key role of *attributed intentions* (versus final outcomes) in determining social sanctions and rewards is also well established (e.g., Falk et al. 2008). Self-image concerns have similar effects, as raising an agent's awareness of his own choices or/and prevailing ethical standards also corresponds to increasing μ. Many experiments indeed document that such manipulations promote fairness and honesty (Batson et al. 1999) and reduce cheating (Beaman et al. 1979), in both performance tests and paid work (Diener and Wallbom 1976; Vallacher and Solodky 1979, Mazar et al. 2008). An even more direct test is provided by Falk (2016), where participants can earn money by inflicting a (real) electric shock on someone else ($a = 0$), or choose not to ($a = 1$). When μ is exogenously increased by exposing subjects to their literal "self-image" –a real-time video feedback of their own face, or a mirror– the likelihood of inflicting the harm decreases by about 25%.

5. Initial self-view (ρ or $\bar{v}$). The model predicts less ethical choices the higher is initial reputation, a set of behaviors that corresponds to what social psychologists term "moral licensing" (for the reputation-rich) and conversely "moral cleansing" (for the reputation-poor). There is ample experimental evidence on these effects in several domains, such as: political correctness (Bradley-Geist et al. 2010; Effron et al. 2009; Merritt et al. 2010; Monin and Miller 2001); selfishness in allocation and consumption choices (Jordan et al. 2011; Khan and Dhar 2006; Mazar and Zhong 2010; Sachdeva et al. 2009); and even dieting (Effron et al. 2012).

6. Seeking or avoiding "the ask". Because image is a "positional good," in fixed total supply $\bar{v}$, any reputational gains of the high type are exactly offset by losses of the low type, leaving on net just the signaling costs expended in the process. On the benefit side, esteem incentives alleviate the self-control problem arising from tempting impulses to behave badly. We thus show that, *ex-ante,* an agent will shun explicit tests of his moral character when μ and/or β are high enough,

$$c\,(1-\beta) < \mu(v_H - \bar{v}), \tag{4}$$

meaning that oversignaling is more of a concern than commitment. When (4) is reversed, he will actively seek moral scrutiny and invest in reputation-sensitive social capital.[6] Both types of strategies are observed in practice, with patterns lining up with the predictions. For help with self-control problems through increased social monitoring, people join religious organizations and peer groups such as Alcoholics Anonymous (Battaglini et al. 2005). Conversely, they tend to avoid situations where social pressure would lead them to be excessively generous. In Della Vigna et al. (2012), for instance, many avoid being home at times when someone soliciting charitable contributions is scheduled to come knock on their door.

A closely related strategy is avoiding even *information* that would provide too explicit a test of one's morality, as when changing sidewalks so as not to walk by a beggar. In Dana et al. (2007) and Grossman and van der Weele (2017), many subjects thus choose not to know whether their choices harm or benefit others. In Exley (2016), they select risky or safe allocations in ways that make inferences about the selfishness of their (anonymous) choices more difficult. Other avoidance strategies include eschewing environments in which sharing is an option (Lazear et al. 2012, Oberholzer-Gee and Eichenberger 2008), or delegating decisions to a principal-biased agent (Hamman et al. 2010, Bartling and Fischbacher 2012). In all these cases, prosocial allocations are significantly less frequent than in identical games that do not allow for such "reputation-jamming" strategies.

2.3 Introducing narratives: exoneration and responsibility

Besides intrinsic moral values and (self-)image concerns, the third key determinant of how people behave –and are judged– are beliefs about the externality e involved in their choices. In Proposition 1, actor and observer share the same belief; whenever they do not, arguments

[6] See Proposition 7 in the Appendix. Dillenberger and Sadowski (2012) offer an alternative formalization of "avoiding the ask," based on temptation preferences rather than signaling concerns.

about what constitutes moral or immoral behavior will come into play.

Definition. A (moral) narrative 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 value e, distributed ex-ante on $[0, 1]$ according to a cdf $F(e)$. We provide concrete examples of the two main classes of such arguments, then in their light discuss the formal components of the definition.

(1) *Absolving narratives* or *excuses* serve to legitimize selfish, short-sighted or even intentionally harmful actions, by providing representations and rationalizations of such acts as consistent with the standards of a moral person. These socially "negative narratives" operate through exculpatory or neutralization strategies (Sykes and Matza 1957) such as: (a) downplaying the harm**;** (b) blaming the victims; (c) denying agency or responsibility; (d) appealing to higher loyalties like religious values or missions that justify hurting others in the name of "a greater good."

Typical of (a) are sanitizing euphemisms such as the military "taking out" people and carrying out "surgical strikes" with "collateral damage"; the framing of a nuclear-reactor accident as a "normal aberration" (Bandura 1999); and describing lies as "a different version of the facts" (Watergate hearings, see Gambino 1973) or, nowadays, as "alternative facts." Extreme uses of (b) include degrading victims as "subhuman," as in the Nazi propaganda against Jews (Levi 1988, Zimbardo 2007) and that of the Hutu government against Tutsis (Yanagizawa-Drott 2014). Common instances of (c) are statements like "we just followed orders" and "I am just doing my job", or underestimating being pivotal, as in the bystander effect (Darley and Latane 1968): "if I don't do it, someone else will." Finally, a vivid example of (d) is the systematic use of narratives and analogies from the Old Testament to support the Indian Removal policy and related atrocities in 19th century America (Keeton 2015).

(2) *Responsibilizing narratives,* on the contrary, create pressure to behave well, by emphasizing how a person's actions impact others, as well as the moral *responsibility* and inferences that result from such agency: making a difference, setting an example or precedent, etc. Examples of such *positive narratives* include: (a) appeals to moral and religious parables, inspiring myths or role models; (b) arguments and cues inducing empathy ("What if it were you?"), making salient the plight of others (identifiable-victim effect) and the personal benefits of good behavior ("You will feel good about yourself"); (c) stressing common identities, such as national and religious brotherhood, sharing the same planet, etc.; (d) appealing to Kantian-like arguments ("What if everyone did the same?") or again invoking some higher moral authority that will pass judgement (God, Adam Smith's "impartial spectator within the breast," "your mother if she could see you," etc.). Many scholars have also argued that oral, written and cinematic stories are essential components of "effective moral education" (Vitz 1990, p. 709).

Discussion. Let us now comment back on our formal representation of narratives as signals 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". Exactly how high or low they have to be to actually alter moral behavior will be determined in equilibrium.

Second, as many of the examples show, stories need not be objectively true to nonetheless powerfully influence a person's behavior and judgment (see, e.g., 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 the wrong conclusions, due to a variety of systematic biases –confusing correlation with causation, framing effects, base-rate neglect, similarity-based reasoning, etc.; (c) unsubstantiated, invented or illogical arguments that nonetheless strike a chord at an intuitive or emotional level, or play into wishful thinking.[7]

Thus, abstracting from the specific channels through which any given narrative may persuade, we focus instead on *why and how people use them,* in a social equilibrium. The essential feature for any *positive* analysis is indeed that these stories or messages "work" –be subjectively perceived by recipients as containing enough of a "grain of truth" to affect their inferences and behaviors.[8] For *normative* conclusions their veracity or falsehood 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 (or philosopher) knows or deems to be "the Right one".[9]

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.[10]

Finally, the examples make clear that the probability an agent will learn of different narratives is likely to be endogenous.[11] It is useful to distinguish two main channels, though in practice they often operate jointly. The first and richer one is the *transmission* channel, where each individual may learn of the narrative by a friend, neighbor or colleague, and can in turn share it with someone else; this is the focus of the next section. The second one is the *production* channel, where an agent obtains the signal through his own costly search for reasons to act morally or selfishly, then decides again whether to disclose (or dwell on, keep in mind, etc.) the arguments he came up with. This is analyzed in the Supplementary Appendix.

[7] On (b), see Tversky and Kahneman (1974), Mullainathan et al. (2008) or Bordalo et al. (2016). On (c), see the *Journal of Economic Literature's* Symposium on Motivated Beliefs: Bénabou and Tirole (2016), Gino et al. (2016) and Golman et al (2016). Examples include, on the negative side,"easy-fix" solutions (tax cuts will pay for themselves), and conspiracy theories. On the positive one, the "Imagine everyone did this" argument, related to Kant's categorical imperative, is ubiquitous. Since one's action will *not* become "universal law", and a bit more pollution "makes no difference", why is it so powerful? Our conjecture is that it may be hard to figure out whether a small externality e (one instance of littering) justifies a small cost c (walking to the next trashcan). The narrative magnifies the salience of both (one envisions dirty cities), facilitating the comparison.

[8] Some of the most successful narratives are even demonstrably wrong: Protocol of the Elders of Zion and other conspiracy theories, pseudo-scientific denials of global warming, and other "alternative facts." Barrera et al. (2017) find that incorrect facts embodied in an effective compelling narrative have a much stronger influence on voting intentions than actual ones, and that correcting the facts does nothing to undo these effects.

[9] If her objective function is, like the agents', linear in the externality, this is very similar to just rescaling the latter's valuations v_H, v_L for the welfare analysis; see Section 4.

[10] For instance: (i) under the model's interpretation in which only truly informative signals can be effective narratives, one adds the Bayesian constraint that $E_F[e] = e_0$; (ii) in the Supplementary Appendix we show how the lower and upper tail moments of F (option values) critically shape the equilibrium moral standard.

[11] The case of exogenous narratives maps directly (normalizing $x \equiv 1$) into Proposition 1, with negative (resp., positive) narratives corresponding to $e \leq e^*$ (resp., $e > e^*$). It also arises when agents receive messages from a "narrative entrepreneur" who wants to induce a fixed behavior. For instance, Glaeser (2005) models politicians who seek to expand their power by systematically broadcasting stories that sow hatred against some minority.

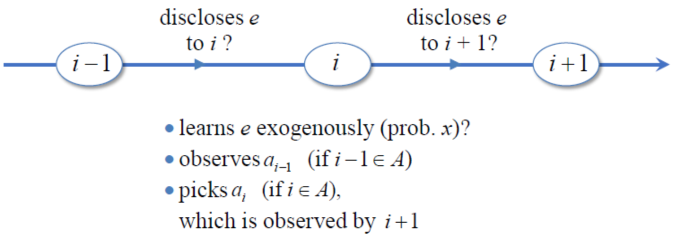
discloses e to i ?
discloses e to $i+1$?
$i-1$
i
$i+1$
• learns e exogenously (prob. x)?
• observes a_{i-1} (if $i-1 \in A$)
• picks a_i (if $i \in A$),
which is observed by $i+1$

Figure 1: Viral Transmission of Narratives

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)

Narratives, by definition, get narrated –passed on from one person to another, thereby potentially exerting considerable influence on a society's moral judgments and actions. We analyze here the different mechanisms through which exculpatory versus responsibilizing 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 *equally 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 promotes others' moral behavior.[12]

3.1 Signaling and disclosure on a stochastic network

1. Setup. There is a countable set of individuals $i \in \mathbb{Z}$, arranged on a line. Each can be of one of two activity types: "Passive" (equivalently, "Principal"), in which case he has no opportunity to choose a moral or immoral action and this is known to his successor $i+1$; or "Active," in which case he does choose $a \in \{0,1\}$, and this action is observed by $i+1$. Equivalently, each agent could have multiple successors, in which case the network is a doubly infinite 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

[12] 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.

successor is active or passive –i.e., exactly who will learn of his behavior, 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.[13] 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:

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

Agents' preferences remain unchanged: a proportion ρ has moral type v_H and the remaining $1-\rho$ type v_L, with $v_H > v_L > 0$, and all active agents share the same reputational concern μ with respect to their audience, which for i is simply his successor $i+1$. As explained above, each individual's moral preference v now logically 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). In particular, if agent i's transmitting e to $i+1$ can be expected to ultimately switch N decisions from 0 to 1 (respectively, 1 to 0), he values this as vNe (respectively, $-vNe$).[14]

The distribution of potential signals or narratives is, for simplicity, taken here to be bimodal: e equals e_- (probability f_-) or e_+ (probability f_+), with $e_- < e^* < e_+$. 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, then chooses 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 contests between opposing rationales, this framework will nonetheless allow us, by averaging across events (episodes, "eras," etc.), to analyze how the conversations, beliefs and behaviors of a society and different subgroups within it are shaped by the competing influences of both types of arguments.

2. Applications. A topical example pertains to *norms of gender relations* in the workplace. Men take actions or say things that affect the welfare of women (e), but they are a priori uncertain (some might say: "have no clue") of whether those will be experienced as innocuous flirting, unwelcome advances or even traumatizing harassment. Various narratives consistent with one view or another circulate, both publicly relayed by the media (probability x) and passed on between people: personal experiences, #metoo testimonies, high-profile cases proved or discredited, polls, movies, cultural stereotypes, etc. Some men genuinely care about not harming women (v_H), others are indifferent or misogynistic (v_L), but all predominantly want

[13] "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.

[14] Since N (computed below) will always be finite, each 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 and his decisions depend only on his role as a *source* of externalities, which he cares about intrinsically. In smaller groups or networks agents would internalize their purely selfish return from strategically affecting $\bar{a}$, but this would be essentially equivalent to renormalizing the v's.

to be seen as being of the first type. The same framework clearly applies to how a dominant national group will "treat," and justify treating, ethnic minorities or immigrants.[15]

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 $a_i = 0$. In both cases, reputation concerns are the same for both moral types but the v_H ones have a stronger influence concern, so they are more inclined to spread positive narratives and refrain from spreading negative ones.[16]

4. Narrations as substitutes or complements. The strength of influence motives also depends on how much further the argument is expected to be spread and affect decisions, giving rise to a *social multiplier.* As the above reasoning suggests, we shall see that negative (absolving, guilt-immunizing, antisocial) narratives are *strategic substitutes*, in that a higher propensity by successors to transmit them makes one more reluctant to invoke them. Conversely, positive (responsibilizing, prosocial) narratives will be *strategic complements,* with individuals' willingnesses to disclose amplifying each other's.

5. Equilibrium. To limit the number of cases, we focus on (stationary) equilibria where:

(1) Whenever an excuse for behaving selfishly is available, the reputation-preservation motive prevails over the influence concern. Thus, upon learning of $e = e_-$, 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. In contrast, the influence effect will be paramount in the propagation of

[15] Even in a "jock" subculture, abusers do not boast that they caused harm and trauma, but instead argue that the act was what the victim "really" wanted, consented to, said no but meant yes, etc. When hurting an outgroup is actually seen as public good for the ingroup ("keep them in their place, teach them a lesson, etc.") only the interpretation of e changes. For the case of multiple audiences with conflicting notions of prosociality, see the Appendix. Groups A and P differing in size could also be accommodated by making the Markov chain (5) asymmetric, and they could also play mirror roles, with P's taking another action that affects the A's.

[16] These tradeoffs are dampened if agents have access not only to "common value" excuses that can be reused by others, but also to "private value" ones (e.g., idiosyncratic shocks to cost c) that cannot. Such an extension of the model would be relatively straightforward.

positive narratives, $e = e_+$.[17]

(2) 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 $a_H(\emptyset) = 0$.[18] This default can be interpreted as the *prevailing social norm,* either "strong" or "weak". 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) we focus on the case where reputational concerns dominate influence ones; the resulting pooling reputation is then $v_D = \bar{v}$. Responsibilizing 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 $a = 1$ is fully revealing) and self-incriminating for the low type. Making use of (6), it follows 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. 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. Therefore,

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

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

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

[17] 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, in contrast, for agents who are small relative to network size, as in our model and the main applications discussed below.

[18] 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.

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 the same as when observing $(a_{i-1}=0, ND)$, resulting in the same expected externality $\hat{e}_{ND}$.

Case 2. Predecessor is passive. When $i-1 \in P$, lack of disclosure means that either he was uninformed or that he censored a signal e_-. This results in:

$$\tilde{e}_{ND} \equiv E\left[e \mid i-1 \in P, ND\right] = \frac{f_- e_- + f_+(1-x)e_+}{f_- + f_+(1-x)} < e_0. \tag{13}$$

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

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 $a_H(\emptyset)=1$ 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:

$$N_-^A = (1-x)\lambda(\rho + N_-^A) \iff N_-^A = \frac{(1-x)\lambda\rho}{1-(1-x)\lambda}. \tag{14}$$

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

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

$$v_H e_-(1+N_-^A) - c \leq \mu(\bar{v} - v_H), \tag{16}$$

$$v_H \tilde{e}_{ND} - c > \mu\left(v_H - v_L\right), \tag{17}$$

with $\tilde{e}_{ND}$ defined by (13). The first one states that, when informed of e_-, even a high-type agent 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 $a=1$ and censor the news that $e=e_-$. Both inequalities show that disclosures of *negative (absolving) narratives are always strategic substitutes:* the higher is N_-^A, the greater will be the social damage from invoking e_-, making i more reluctant to do so (requiring a higher reputational payoff).

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 $a=1$ and reveal himself rather than $a=0$, which given the unavailability of excuses would misidentify him as a low type. This requirement is more stringent when the "silent" predecessor was a passive agent $i \in P$ than an active one, since we saw that nondisclosure leads to lower inferences about e in the first case relative to the second: that is why the expected externality involved is $\tilde{e}_{ND} < e_0$ rather than $\hat{e}_{ND} > e_0$. Finally, together with (13), the condition shows that an equilibrium with $a_H(\emptyset) = 1$ requires that the prior f^+ be high enough, which is quite intuitive.

Proposition 2 (**morality as the default behavior**). *When (15)-(17) hold, they define 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 (14); 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 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.[19]

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 responsibilizing 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. The reason he does so is that $i + 1$ could turn out to be a passive agent (probability $1 - \lambda$) and thus unable to signal e_+ through his actions; being given the actual narrative will allow him 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,

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

and thus again *censors* positive narratives e_+. As before, both active types pass on negative ones, e_-. Given these action and communication strategies,

$$x_-^P = x + (1-x)(1-\lambda)x_-^A, \quad x_-^A = x + (1-x)\lambda x_-^A, \tag{18}$$

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

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.4)-(A.5) in the Appendix). The "influence factors" or social multipliers are now $N_-^A = 0$ for all agents in the case of e_- (as it will change no behavior), while for e_+ they are

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

$$N_+^A = (1-x)\left[\lambda\rho(1+N_+^A) + (1-\lambda)N_+^P\right], \tag{21}$$

for passive and active agents (of either moral type), respectively. The solutions to this linear system are given by (A.7)-(A.8) in the Appendix.

Consider now the updating. As before, any 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":

$$\hat{v}_{ND} = \frac{\rho(1-\bar{x}_A)v_H + (1-\rho)[1-f_-x_-^A]v_L}{\rho(1-\bar{x}_A) + (1-\rho)\left[1-f_-x_-^A\right]} \in (v_L, \bar{v}). \tag{22}$$

Indeed, i 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

$$\hat{e}_{ND} \equiv E\left[e \mid a_{i-1} = 0, ND\right] = \frac{f_-(1-x_-^A)]e_- + f_+(1-\rho x_+^A)e_+}{f_-(1-x_-^A) + (1-f_-)(1-\rho x_+^A)} > e_0. \tag{23}$$

If the "silent" predecessor $i-1$ was a passive agent, the inference is the same $\tilde{e}_{ND} < e_0$ as before, given by (13). The complete conditions for an equilibrium with $a_H(\emptyset) = 0$ are then

$$v_H e_- N_-^A = 0 < \mu(\bar{v} - \hat{v}_{ND}), \tag{24}$$

$$v_L e_+ N_+^A < \mu\left(\hat{v}_{ND} - v_L\right), \tag{25}$$

$$c - v_H\hat{e}_{ND} \geq \mu(v_H - \hat{v}_{ND}), \tag{26}$$

where $\hat{v}_{ND}$ is defined by (22). The first one verifies trivially that, when informed of an excuse e_-, active agents will always share it (and choose $a = 0$), since it is valuable from a reputational point of view and has no adverse spillover onto followers' behavior.

The second condition states that, when learning e_+, a low type agent prefers to *keep quiet* about it, in order to 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, in contrast to negative ones, sharing decisions are *strategic complements.* The more others tend to pass them on (the higher is 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."

The third condition states that, absent any narrative, a high type indeed chooses $a_H(\emptyset) = 0$ rather than deviating to $a = 1$, which would clearly identify him but not persuade $i+1$ that $e = e_+$, since if he knew that he should have disclosed it. In contrast to the previous ($a_H(\emptyset) = 1$) type of equilibrium, the expected externality is now $\hat{e}_{ND} > e_0$ rather than $\tilde{e}_{ND} < e_0$, 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). Referring to (23), finally, shows that an $a_H(\emptyset) = 0$ equilibrium requires that the prior f^+ not be too high, which is again intuitive.

Proposition 3 (selfishness as the default behavior). *When (25)- (26) hold, they define 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.*

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.7) and (A.8). 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 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 $i+1$ they are 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 different types of equilibria and social norms are associated to endogenously different circulating narratives. In the "moral" equilibrium ($a_H(\emptyset) = 1$), doing the right thing (e.g., respect toward women) *"goes without saying,"* while deviating requires a justification, so negative narratives are the ones that will get passed on (when they occur) and affect

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

behavior. In the "amoral" equilibrium ($a_H(\emptyset) = 0$) self-indulgence 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.[21]

Second, even though the two types of equilibria involve radically different norms and outcomes, Propositions 2-3 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 (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 4 (**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 men and women (say) interact mostly within segregated pools (high λ), very different types of narratives will circulate within each one, with men mostly *sharing rationalizations* for their behavior, which will be worse on average than under integration, and women mostly reasons for why it is *inexcusable.*

Notably, the polarization-minimizing pattern is not a deterministic alternation of agents but a *random* one, $\lambda = 1/2$ –or, equivalently, a *tree network* in which each individual's audience is a 50-50 mix of A's and P's. In contrast, for $\lambda = 0$ 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 *women* who are more likely to hear (from a man) of a narrative excusing controversial male behavior. Similarly, when responsibilizing e_+ narratives circulate, if $\lambda = 0$ (or is low), *men* are more exposed to them, since any woman who knows it will relay it, whereas only high-morality men will disclose it to a woman.

Empirical implications. Whether for gender, ethnicity or income, positive correlation ($\lambda \geq 1/2$) is by far the most relevant scenario, leading to differences in exposure and beliefs of the intuitive rather than "paradoxical" type, for both e_- (excuses) and e_+ (say, #metoo). Assortative social communication arises from reasons taken here as exogenous (homophily, targeted messages, power relationships) 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.

[21] Both equilibria may coexist for some range of parameters (e.g., prior distribution F), with Pareto-dominance having little bite for selection. First, passive agents naturally prefer more moral outcomes. Second, in many cases each actor is himself impacted by the externalities generated by others: pollution, tax evasion, how women are treated at work, etc. Depending on how large e is, this may or may not dominate the fact that, from the sole point of view of their own actions, both H and L types prefer to be in an equilibrium with more relaxed moral standards. Third, as in an overlapping-generations model, coordination on a particular equilibrium requires agreement between an infinite chain of individuals who do not directly communicate, or even coexist.

[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 4 hold in the strict sense, including for the global maximum at $\lambda = 1$.

3.5 Enriching the communication channels

We identified the reputation and influence motives as the two key drivers of moral discourse, then showed how network structure shapes their interplay, diffusion strengths, and the resulting beliefs and norms of a society or subgroups. In Section 4 below, we expand *influence-motivated* communication, now focusing on a single (P, A) dyad and allowing P to have both less rigid preferences and a richer menu of moral-persuasion devices.

In the Supplementary Appendix, we conversely expand *reputation-motivated* communication: focusing on a single (A, P) dyad, we allow A to engage in his own search for reasons for behaving one way or the other. The mere fact that someone has an excuse then suggests that he perhaps sought one, or more generally "looked into the question" of where e lies in $[0, 1]$; this, in turn affects the audience's (P) view of his morality.[23] Formally, the probability x that a narrative from $F(e)$ is *generated* in the first place becomes endogenous and type-dependent. We show that whether v_L or v_H searches more, and thus "how strong" excuses must be to be deemed acceptable –an endogenous *moral standard* $\hat{e}$ such that invoking $e > \hat{e}$ would be worse than offering no justification –hinges not just on the prior mean $E_F[e]$ but, critically, on the *tail uncertainty* (option values) in $F(e)$. In the process, we establish new results on ranking probability distributions according to upper or lower conditional moments.

4 Narratives Versus Imperatives

One actor P is a principal (she) whose only decision is how to communicate with an actor or agent A (he), who will in turn take the pro- or antisocial action. The principal can be thought of as an ex-ante incarnation of the individual, a parent, society, or religious leaders. At her disposal lie several routes of persuasion.

1. Forms of influence. A narrative is again an argument pertaining to the parameters of the agent's decision: externality, cost, or visibility of behavior. In contrast, an imperative is a direct command to behave in a certain way –say, to do $a = 1$; it does not consider motives, only the decision itself. Imperatives typically take the form of succinct, broad precepts, such as the Ten Commandments, and they relate closely to rule-based moral reasoning, put forward by Immanuel Kant. Consequentialist versus deontological normative theories differ precisely in that, for the former, only the ends (life, happiness, welfare) conceivably justify decisions, whereas the latter postulates categorical demands or prohibitions of actions, no matter how (un)desirable the implications.[24] Both strategies are commonly used to instill ethical behavior, and found to be effective in experiments on moral persuasion (e.g., Dal Bó and Dal Bó 2014).

[23] In an intrapersonal, self-signaling context, the search for absolving narratives can also be interpreted as a form of motivated moral reasoning (Ditto et al. 2009).

[24] The discussion about these two main lines of thought in moral philosophy is, of course, more involved (see, e.g., Alexander and Moore 2015). For example, imperatives as they will be modeled here are not unconditionally justified, in contrast to pure deontological principles. Our notion is more akin to so-called "rule consequentialism," which understands an act as morally wrong if it is forbidden by precepts that are themselves justified in terms of their general consequences; see also Kranz (2010). Similarly, philosophers (starting with John Stuart Mill (2002) and more recently Hare, 1993, and Cummiskey, 1996) have suggested a *teleological* reading of the categorical imperative, as a means to produce the best overall outcome.

2. Understanding imperatives. A first role of imperatives is as broad rules-of-thumb or "moral heuristics" (Sunstein 2005) that work well in most cases, but may malfunction in more unusual ones. Because they economize on information and cognitive costs, imperatives provide quick, instinctive guides to decisions in unfamiliar contexts where moral consequences may be complex to figure out. Since they embody only coarse information, on the other hand, they will sometimes induce moral mistakes, at least from a utilitarian point of view.

More puzzling are situations in which the agent is fully cognizant of consequences, yet makes a considered (not instinctive) decision to ignore them in favor of an overriding imperative. In Falk and Szech (2017), for instance, the 18% percent of subjects who act non-consequentially understand, as verified by elicited beliefs, that the their choice has zero chance of having any impact. This inherent rigidity of imperatives points to their second role, namely commitment, which arises from the common tendency of morality judgements to be self-serving, as amply documented in the literature on "moral wiggle room" (e.g., Dana, et al. 2007). This is particularly relevant when self-control is weak, so that even the individual himself may benefit (ex-ante) from imperatives that do not simply substitute for missing information, but also command to ignore any that might be available.

This role of imperatives as "cognitive straightjackets" designed to restrict moral wiggle room can be sustained in two complementary ways. First, it may be directly encoded into strong, visceral preferences (repugnance, compulsiveness) that will trump arguments of reason; such preferences are then non-consequentialist at the individual level, though they may be in terms of evolutionary fitness. Alternatively, it can be sustained by purely utilitarian individuals as a self-enforcing "personal rule" (Bénabou and Tirole 2004). When excuses and rationalizations are too easily "made up" in ambiguous situations, a rule of disallowing even genuine evidence that, "this time," following self-interest will make no difference to others, or that harming someone is required for a greater good, becomes a signaling device allowing stronger moral types to distinguish themselves from weaker ones.[25]

The formal analysis will confirm the importance of the factors foreshadowed above, on both the principal's and the agent's side, as conducive to the use of imperatives.[26] First, imperatives are effective only if issued by trusted principals, whereas everyone can use the narrative route to attempt persuasion. Second, their coarse and cheap-talk nature economizes on cognitive costs and makes them less fragile to misinterpretation (whether accidental or motivated) than narratives. Third, because they focus on the decision itself without allowing for fine contingencies, they entail some costly "moral rigidity" in choices. This also allows for commitment, however, making them potentially valuable to deal with temptations. Fourth, by pooling states in which the agent would be reluctant to behave in the desired way with others where he is willing, imperatives allow the principal to induce broader compliance.

[25] We model in the Appendix a related phenomenon: merely *questioning* an imperative –giving thought to reasons why an exception might be warranted– can be a bad sign of one's morality, and thus "forbidden" or "taboo" in equilibrium.

[26] Kant formulated his categorical imperative from both perspectives; agent (*"Act only in accordance with that maxim through which you can at the same time will that it become a universal law."* (Kant, 1785, 4:421)) as well as principal (*"The Idea of the will of every rational being as a will that legislates universal law"* (Kant, 1785, 4:432)), i.e., both in terms of a universal law giver as well as universal law followers (see Johnson, 2014).

4.1 Modeling imperatives

There is a principal (she) who learns a narrative drawn according to a continuous $F(e)$ on $[0,1]$, and an agent (he) who does not and will choose an action $a=0,1$. There may also be a passive audience forming an image of the agent, which he cares about, though here that is not essential (μ could be zero). The prior mean e_0 is below e^*, so that the agent will not behave prosocially unless prompted by some communication from the principal. The situation is thus similar to that between a passive agent ($i \in P$) with pure influence concerns communicating with an active successor ($i+1 \in A$) in Section 3, but for two differences. First, the principal no longer wants the agent to unconditionally choose $a=1$: her preferred decision depends on the value of the externality, e, in a way that can be more or less congruent with the agent's preferences. Second, besides sharing her signal or narrative she can also, or instead, issue an imperative; in Section 3 the latter would not have been credible, being always state-independent.

Let us denote by $U^A(e)$ and $U^P(e)$, respectively, the moral agent's and the principal's net returns from his choosing $a=1$ rather than $a=0$ (for the low type, $a=0$ is still a dominant strategy). For instance, suppose that agents have the preferences used so far,

$$U^A(e) = v_H e - c + \mu(v_H - \bar{v}) = v_H(e - e^*) \tag{27}$$

where $c = c_0/\beta$ is the cost perceived at the moment of choice, whereas the principal internalizes spillovers on a larger scope (own private benefits, whole population vs. in-group focus, etc.), as well as any "internalities" arising from imperfect ($\beta \leq 1$) self-control. Thus

$$U^P(e) = E_{\tilde{v}}[(we + \tilde{v}e - c_0)a(\tilde{v})] \equiv \rho\left(w + v_H)(e - e^P\right), \tag{28}$$

where $w \geq 0$ is the extra value she attaches to the agent's moral conduct beyond his *ex-ante* welfare, and $e^P \equiv c_0/(w+v_H)$ her indifference point. The gap with the decision threshold $e^* = [c_0/\beta - \mu(v_H - \bar{v})]/v_H$ is larger, generating a greater role for imperatives and/or narratives, the higher is w and the more *present-biased* or less *image-conscious* the agent is. The case $w=0$ corresponds to a sophisticated individual's ex-ante self (or parents maximizing their child's welfare), that of $w=1$ to a utilitarian social planner, and that of $w=+\infty$ to a passive actor wanting to promote the action $a=1$ without any empathy for the agent.

More generally, we will simply assume that U^A and U^P are both *affine functions of* e, with indifference points defined by $U^A(e^A) = U^P(e^P) \equiv 0$ such that $e^P < e^A = e^*$, meaning that the principal favors $a=1$ over a larger set of values than the high-type agent. Fixing e^*, we will identify e^P with the degree of congruence between them and assume that the agent knows $U^P(\cdot)$ –that is, how trustworthy the principal is.[27]

4.2 Coarse versus noisy communication

Absent the possibility of imperatives, the natural equilibrium would be for the principal to communicate all narratives $e > e^*$ and say nothing otherwise, a silence which a rational agent would then correctly interpret as meaning that $e \leq e^*$. Suppose now that when the principal

[27] The case $e^P \geq e^*$ is straightforward: whenever $e \geq e^P$ the principal will issue a credible imperative, and when $e < e^*$ he will keep silent and the agent will choose $a=0$, since $E\left[e|e < e^P\right] \leq e_0 < e^*$.

tries to convey an argument $e > e^*$, there is some small probability $1 - \xi$ that the agent does not receive the message –did not hear it, was not paying attention, cannot make sense of it except as uninformative random noise, or even interprets it the wrong way. In such cases his belief will remain e_0 or may even decrease, leading him to mistakenly choose $a = 0$.

Issuing an imperative of the form "do $a = 1$" without going into reasons is a clearer, much less complex message, not subject to miscommunication. On the other hand, for it to be operative in equilibrium, one must have:

(a) Incentive compatibility: anticipating obedience, the principal orders $a = 1$ if and only if $U^P(e) \geq 0$, or $e \geq e^P$.

(b) Persuasiveness: the (high type) agent obeys the imperative, picking $a = 1$ when told to do so. This requires that

$$\mathcal{M}^+(e^P) \equiv E[e \mid e \geq e^P] > e^*. \tag{29}$$

When (29) holds, it is indeed optimal for the principal to issue imperatives according to (a), and for the agent to follow them, as in (b). This strategy yields payoff $U^P(e)$ in all those states, whereas the "argumentative" strategy of disclosing e yields only $\xi U^P(e)$, and this only for states $e > e^* > e^P$. Provided that $\xi < 1$, (a)-(b) is also the unique equilibrium under the Pareto selection criterion (applied here sequentially to the principal and then the agent). By contrast, there is no equilibrium with an imperative when (29) fails. The principal uses narratives instead, when she has them, and equilibrium compliance is, on average, only $\xi[1-F(e^*)] < 1-F(e^P)$. Condition (29) also delivers comparative statics on the factors favoring the emergence of imperatives.

1. Congruence. As e^P increases so does $\mathcal{M}^+(e^P)$, making the inequality more likely to hold. To convince the agent that she is standing for their interests, the principal thus cannot be too much of an unconditional advocate for pro-social actions (in the weighted-utility illustration of $U^P(\cdot), w$ should not be too high). Principals who are too dogmatic about what is the "right thing to do" will not be listened to.

2. Perceived soundness of judgment. Suppose that a principal P_1 has access to more accurate (or more persuasive) narratives than another one, $P_2:$ formally, the induced distribution of posterior beliefs $F_1(e)$ second-order stochastically dominated by $F_2(e)$. By Lemma 1 in the Supplementary Appendix, it follows that $\mathcal{M}^+_{F_1}(e^P) > \mathcal{M}^+_{F_2}(e^P)$ as long as $F_1(e^P) \leq F_2(e^P)$, i.e. as long as P_1 also has more "positive" priors (or not too worse ones) about the desirability of the agent's contribution. Under that condition, being perceived as better informed confers greater "moral standing" to a principal, allowing her to more credibly issue imperatives: (29) becomes more likely to hold. Narratives, in contrast, can be spread by anyone who has it.[28]

3. Large expected externalities. Suppose that the distribution of e increases uniformly with a shift parameter θ : it has cdf $F(e - \theta)$, and mean $e_0 + \theta$. Assuming that the hazard rate $f/[1 - F]$ is increasing, we have for all $\theta_1 < \theta_2$:[29]

[28] At least when they consist of hard information. When they are messages that exploit salience effects, similarity-based reasoning, logical fallacies with emotional appeal, etc. as also discussed in Section 2.3, this may require particular "talents" of persuasion. Some of these same talents can also be useful in making imperatives credible (e.g., looking authoritative, trustworthy, benevolent, etc.).

[29] Note that $\mathcal{M}^+(e^P, \theta) = \theta + M^+(e^P - \theta)$, and recall that $(\mathcal{M}^+)' \in (0, 1)$ under the hazard-rate condition. Larger externalities in the more general FOSD sense, on the other hand, need not always increase $\mathcal{M}^+$.

$$\mathcal{M}^+(e^P,\theta_1) \geq e^* \Longrightarrow \mathcal{M}^+(e^P,\theta_2) \geq e^*.$$

Proposition 5 (clarity vs. credibility). *Suppose that there is at least a slight probability of miscommunication of any narrative. Then:*

1. *There is a unique (Pareto-dominant) equilibrium: if $\mathcal{M}^+(e^P) > e^*$, the principal issues an imperative whenever $e \geq e^P$ and does not communicate otherwise; if $\mathcal{M}^+(e^P) \leq e^*$, she discloses her narrative whenever $e > e^*$ and does not communicate otherwise.*
2. *The use of imperatives is more likely for a principal who is perceived as having greater moral authority, in the sense that her interests are more congruent with those of the agents, that she is better informed (and not too pessimistic) about externalities from their actions and/or these externalities are likely to be (uniformly) more important a priori.*

That sufficient congruence is a prerequisite for imperatives accords well with the fact they are much more common and effective in parent-child relations than between loosely related interaction partners.[30] Likewise, the fact that moral authority or “wisdom” is a precondition sheds light on why religious leaders can rely on them much more than politicians, who instead must usually appeal to narratives. Finally, imperatives being more likely when stakes (externalities) are high fits well with the observation that the strongest and most universal ones pertain to issues of life, health and reproduction.

4. Combining narratives and imperatives. Morals systems, religions and educators often blend the two types of arguments, as in fables where someone stole or lied then came to regret it, followed by a generalization to “thou shalt not steal/lie.” A general treatment lies outside the scope of this paper; we simply outline here an example showing how, when congruence is too low for a credible decree, the principal may start with some narrative(s) that raise her authority enough that an imperative then becomes effective.

Let $\hat{e} < e^*$ be defined by $\mathcal{M}^+(\hat{e}) = e^*$, and suppose that $e^P \leq \hat{e}$, so that (29) fails. Assume that the principal receives (with some probability) a coarse signal, which can be disclosed without risk of misunderstanding and raises the posterior to $e' > \hat{e}$. In a second stage (or simultaneously), she learns the actual e, but that more precise narrative is harder to communicate –subject to an error rate $1-\xi$, as before. When the coarse narrative is received it will be disclosed, and this in turn renders credible issuing the imperative “do $a = 1$” for all values $e \geq e^P$, whereas on its own it would fail.

4.3 The value of flexibility

Suppose now that the agent also has or can obtain private signals about the potential externality, of a type that is only relevant when combined with information disclosed by the principal. This could be some complementary data or information search, or equivalently some thought process through which the principal’s stated narrative is combined with the agent’s own experience.

By contrast, we assume that an imperative does not trigger such thinking or information retrieval. For instance, providing arguments to an agent as to why he should do something

[30] Empirically, parents not only place high value on the utility of their children, they are also similar in terms of their preferences (Dohmen et al., 2012), whether due to genetic or cultural transmission.

may lead him to think more about it and perhaps find valid counter-arguments (with which the principal would agree), whereas a trusted principal telling him to "do it because I say so" will not lead to any further information being brought in. Hence, again, a tradeoff.

Formally, suppose that when provided with narrative e, the agent arrives at a final assessment of the externality ε that is distributed according to some differentiable function $H(\varepsilon|e)$, with $E(\varepsilon|e) = e$ and $H(e^*|e) < 1$ for all e, such that: (a) an increase in e shifts the distribution of ε to the right in the sense of the monotone-likelihood-ratio property, i.e. $H(\varepsilon|e_2)/H(\varepsilon|e_1)$ is increasing in ε if $e_1 < e_2$; (b) ε is a sufficient statistic for (ε, e), implying in particular that the principal also wants to evaluate final payoffs, and thus the agent's choices, according to the posterior beliefs ε. From (b), U^P and U^A depend on ε, not on e. Let us denote by

$$V^P(e) \equiv \int_{e^*}^{1} U^P(\varepsilon) dH(\varepsilon|e)$$

the principal's welfare under a strategy of disclosing her narrative e, and look for conditions under which she prefers (in equilibrium) to instead issue an imperative to "do $a = 1$" over some subset of states, denoted I; obedience by the agent requires that $E[e|e \in I] \geq e^*$. The advantage of the narrative strategy is its *flexibility*, valued by both parties: whenever the principal's signal would call for action, $e > e^P$, but the moral agent's information combined with it leads to a low final posterior $\varepsilon < e^P$, both will concur that he should choose $a = 0$, whereas under an (effective) imperative he would have chosen $a = 1$. Equilibrium behavior therefore requires that

$$\Delta(e) \equiv \int_{0}^{e^*} U^P(\varepsilon) dH(\varepsilon|e) \geq 0 \quad \text{for all } e \in I, \tag{30}$$

and conversely $\Delta(e) \leq 0$ for $e \notin I$. Thus, $-\Delta(e)$ is the "value of information" to the principal, conditional on her signal e. Note that (30) is never satisfied at $e = e^P$, since

$$\Delta(e^P) + \int_{e^*}^{1} U^P(\varepsilon) dH(\varepsilon|e^P) = \int_{0}^{1} U^P(\varepsilon) dH(\varepsilon|e^P) = U^P(e^P) = 0, \tag{31}$$

where the last equality results from the linearity of U^P; thus, $\Delta(e^P) < 0$. Under the monotone-likelihood-ratio property, moreover, if $\Delta(e_1) \geq 0$ for some e_1 then $\Delta(e_2) > 0$ for all $e_2 > e_1$ (see the Appendix). Therefore, I is of the form $I = \left(e^\dagger, 1\right]$, with $e^\dagger > e^P$, and an imperative exists –is issued in equilibrium for some values of e– if and only if

$$\mathcal{M}^+(e^\dagger) \geq e^*, \quad \text{where} \quad \Delta(e^\dagger) \equiv 0. \tag{32}$$

1. Congruence. Suppose that congruence increases uniformly, in the sense that $U^P(e)$ shifts down for all e (say, in the weighted-utility cases, w decreases). This causes $e^\dagger$ to rise, so (32) becomes more likely to hold and the principal more willing to delegate decision-making to the agent. Effective imperatives thus require a minimum amount of "trust" by the agent, but as the two parties' interests become even further aligned, the principal finds sharing narratives (when available) increasingly valuable relative to issuing the imperative, and the frequency $1 - F(e^\dagger)$

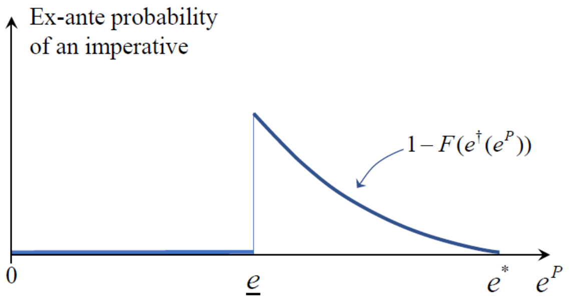

Figure 2: Impact of Congruence on the Use of Imperatives

of the latter decreases toward 0; see Figure 2.[31]

2. Self-control. The analysis, applied with the benchmark preferences (27)-(28), reveals a similarly non-monotonic impact of self-control. As β decreases (with c_0 constant), a rigid personal rule becomes more likely to emerge: e^* increases, so $I = (e^{\dagger}, 1]$ expands. At some point, however, willpower becomes weak enough that the obedience condition $\mathcal{M}^+[e^{\dagger}(e^*)] \geq e^*$ fails, and having a strong narrative $e > e^*$ becomes indispensable. On Figure 2, decreases in β shift the points $\underline{e}$ and e^* to the right, which is equivalent to reducing e^P.

Proposition 6 (congruence and flexibility). *Suppose that the agent can use private information to refine the principal's narrative, so that imperatives have a cost in terms of flexibility. Define $e^{\dagger} > e^P$ by $\Delta(e^{\dagger}) \equiv 0$, as in (30).*

1. *Imperatives are used in equilibrium if and only if $\mathcal{M}^+(e^{\dagger}) \geq e^*$. In that case an imperative is issued whenever $e \geq e^{\dagger}$, whereas for $e < e^{\dagger}$ the principal discloses her narrative.*

2. *The probability of an imperative being used is hump-shaped in congruence: zero below some minimum level, then positive but decreasing back to zero as the alignment of interests increases further. The effect of self control on imperatives is similarly hump-shaped.*

Refraining from questioning an imperative. From a deontological perspective, imperatives must be obeyed irrespective of consequences, that is, of resulting costs and benefits. Even the very act of *questioning* the imperative, and not only violating it, can therefore be a dangerous path. In Bénabou and Tirole (2011a), merely thinking about the price attached to a taboo or a "repugnant" transaction damages the individual's self-worth, even if the deal is not concluded in the end. In the Appendix, we add the idea that the agent could challenge an imperative issued by a principal. "Calculating" individuals who question the imperative, however, may reveal themselves as persons of mediocre moral standing, even when they end up behaving prosocially. If this loss in reputation or self-esteem is sufficient, they will not question the rule or edict, thus engaging in information avoidance to mimic individuals with high enough moral standards as to not even give it a second's thought.

[31] The figure is drawn for parallel shifts of U^P, which can thus be indexed by the intercept e^P only. The threshold $\underline{e}^P$ is defined by $\mathcal{M}^+(e^{\dagger}(\underline{e}^P)) = e^*$.

5 Concluding Remarks

We developed a flexible framework for analyzing moral behavior and discourse. Besides known standard factors such as intrinsic preferences, self-control and social- or self-image concerns, it incorporates a critical new one, namely the generation, use and circulation of *arguments* –such as narratives and imperatives– about the moral consequences of one's actions. The model also helps to organize and interpret many empirical findings, and generates new predictions.

Many issues remain to be explored. First, we modeled narratives as *acting as* hard signals about social or/and private payoffs, while stressing that in practice they may or may not, upon closer inspection, have real informational content or be logically coherent. Put differently, we took as a primitive some class of arguments that "work" in persuading agents about social externalities, and focused on analyzing how people will then *search* for them, *invoke* them, *repeat* them, and *judge* those who do so. What makes many "content-free" narratives work involves important elements of heuristic and/or motivated thinking; existing models of these could easily be combined with the present one.

Another potential extension is to conflicting narratives. In the model, at each point in time a single (but potentially different) moral argument is circulating, which agents may be exposed to and strategically use; in reality, interest groups and "narrative entrepreneurs" will simultaneously offer very different rationales for what is right or wrong. What factors then make one story more compelling than the other? Interesting applications would include politics (propaganda, fake news), identity conflicts (in-group/out-group narratives), and the optimal mix of narratives and imperatives in the design of religious and other doctrines.

Arguments about what constitutes a moral act also extend beyond beliefs about its consequences. We think that any serious model of morality must consider externalities –causing or avoiding harm to others– but acknowledge that other notions may be relevant as well. Haidt (2007), for instance, criticizes the reduction to the fairness-harm conception and suggests the inclusion of phenomena such as loyalty, authority, and purity. These notions can, in large part, already be mapped to our model through the (self) signaling of personal values and the in-group they extend to, but working out more specific applications seems worthwhile.

Differing social preferences even under full information (alternatively, heterogenous priors) constitute another potential source of disagreement, often relevant for "hot" societal issues such as religion, abortion, immigration, etc. When audiences disagree on what constitutes a negative versus positive externality, social image becomes multidimensional. We explained how the model can capture the net effects of these concerns, given how their relative importance to different individuals reflects the degree of assortative matching in society. The latter is ultimately endogenous, however, so it would be interesting to analyze group or network formation jointly with social signaling and narrative transmission.

Another natural direction is experiments. On the narratives side, Foester and der Weele (2018b) provide evidence supporting the implications, shared by their and our model, of how reputation and influence incentives shape strategic communication in a dyad. On the imperatives side, we analyze in Bénabou et al. (2018) how different methodologies formally compare in eliciting true moral preferences, and what this implies for how to interpret Kantian-like refusals of tradeoffs that appear deontologically rather than consequentially motivated.

Appendix: Main Proofs

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, is 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 $(v_1, v_2) \in \{0, v_L, v_H\}^2$.

Seeking or avoiding moral choices We examine here when, on average (or, behind the veil of ignorance about one's type), it is better to face a restricted choice $\{a = 0\}$, generating utility $\mu\bar{v}$, or a moral decision $a \in \{0, 1\}$ in which (social or self) image $\mu\hat{v}(a)$ is also at stake. In the latter case, ex-ante utility is, with $c_0 \equiv \beta c$,[32]

$$U_0 \equiv \rho U_H + (1-\rho)U_L = E[(ve - c_0)a + \mu\hat{v}(a)] = E[(ve - \beta c)a] + \mu\bar{v}, \qquad \text{(A.1)}$$

by the martingale property of beliefs. Comparing (A.1) to the ex-post $U = (ve - c)\,a + \mu\hat{v}(a)$ given by (1) shows that: (i) oversignaling occurs when $e > e^*$, i.e. $v_H e > c + \mu(v_H - \bar{v})$, but $v_H e < \beta c$; (ii) conversely, there is undersignaling when $e \leq e^*$ but $v_H e > \beta c$.

Proposition 7 (avoiding or seeking the ask). *Ex ante, the agent will:*

1. *"Avoid the ask", out of concern for oversignaling, when $e^* < e < \beta c / v_H$. This occurs for a nonempty range of externalities when μ or β is high enough:*

$$c\,(1 - \beta) < \mu(v_H - \bar{v}). \qquad \text{(A.2)}$$

2. *"Seek out the ask" (even at some cost), using image as a means of commitment, when*

$$\max\{\beta c / v_H, e^*\} < e < c/v_{H.} \qquad \text{(A.3)}$$

[32] We have assumed that the warm-glow utility ve is subject to hyperbolic discounting, as it presumably lingers longer than the perceived cost. We could have made the opposite assumption, in which case moral behavior would require $(v_H e/\beta - c) + \mu(v_H - \bar{v}) > 0$. The comparative statics would remain unchanged, but now there would always be oversignaling, so agents would systematically try to avoid moral-choice situations, or decrease their visibility and salience, μ. Since one commonly sees people seeking out visible opportunities to demonstrate their goodness (making named donations, joining NGO's), as well as others who "avoid the ask," we focus on the parametrization that allows for both types of behaviors.

3. *Be concerned about undersignaling, and thus seek greater exposure (increasing μ), whenever (A.2) is reversed.*

Proof of Proposition 2 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 3 We first solve the system (19) to obtain

$$x_+^P = [1-(1-x)\rho(2\lambda-1)]\left(\frac{x}{Z}\right), \tag{A.4}$$

$$x_+^A = [1-(1-x)(2\lambda-1)]\left(\frac{x}{Z}\right). \tag{A.5}$$

where

$$\begin{aligned} Z &\equiv [1-(1-x)\lambda][(1-(1-x)\rho\lambda] - (1-x)^2(1-\lambda)^2\rho \\ &= 1-(1-x)(1+\rho)\lambda + (1-x)^2\rho(2\lambda-1) \end{aligned} \tag{A.6}$$

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

$$N_A^+ = \frac{(1-x)\rho(\lambda-(2\lambda-1)(1-x))}{Z} \tag{A.7}$$

$$N_P^+ = \frac{(1-x)(1-\lambda)\rho}{Z}. \tag{A.8}$$

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

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

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

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

Turning now to the last result in the proposition, each active agent now has probability x_+^A (given by (19)) 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

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

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

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

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

$$\begin{aligned}
A &= ((4\lambda-2)(\lambda(\rho+1)-2\rho(2\lambda-1)) - (2\lambda-1)(\lambda(\rho+1)-2\rho(2\lambda-1)) + \rho(2\lambda-1)(2\lambda-2)) \\
&= \lambda(1-\rho)(2\lambda-1), \\
B &= (4\lambda-2)(\rho(2\lambda-1)-\lambda(\rho+1)+1) = -2(1-\rho)\left(2\lambda^2-3\lambda+1\right), \\
C &= -(2\lambda-2)(\rho(2\lambda-1)-\lambda(\rho+1)+1) = 2(1-\rho)(1-\lambda)^2
\end{aligned}$$

It is monotonic in x, since $P'(x)/[2(1-\rho)] = \lambda(2\lambda-1)x - \left(2\lambda^2-3\lambda+1\right) = (2\lambda-1)[1+\lambda(1-x)]$. Moreover, $P(0) = C > 0$ and $P(1)/(1-\rho) = \lambda(2\lambda-1) - 2\left(2\lambda^2-3\lambda+1\right) + 2(1-\lambda)^2 = \lambda > 0$, therefore $P(x) > 0$ for all $x \in [0,1]$, implying the desired result. ■

Proof of Proposition 4 For negative signals, x_-^P and x_-^A are given by (18), 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.4)-(A.5), we have:

$$\frac{x_+^P}{x_+^A} = \frac{1-(1-x)\rho(2\lambda-1)}{1-(1-x)(2\lambda-1)}, \tag{A.9}$$

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 Proposition 5 Existence is obvious. For Pareto dominance, note that for any "imperative" message m that induces $a_H > 0$, it must be that $E[e|m] \geq e^*$. If the principal does go the imperative route she will then pick an m that induces the highest a_H, so without loss of generality we can focus on a single such message and write her problem as:

$$V^P(e) = \max\left\{\mathbf{1}_{\{e \geq e^*\}}\xi,\ a_H(m)\right\} \times U^P(e).$$

All types in (e^P, e^*) will therefore prefer to issue m. Furthermore, $\xi \leq a_H(m)$, otherwise all types in $[e^P, 1]$ would disclose their narrative rather than issue m, implying that $E[e|m] \leq e^P < e^*$. Thus there is no loss of generality in assuming that all types in $[e^P, 1]$ issue issue imperative m. If $a_H(m) < 1$, such and equilibrium is dominated by the one described in the text. ■

Proof of Proposition 6 Here again existence is obvious, and since the imperative successfully induces $a_H = 1$, the principal gets her highest possible utility, implying Pareto dominance as in the proof of Proposition 5. (Recall that the payoff when disclosing a narrative is equilibrium-independent).

The only claims remaining to prove are equation (31) and the cutoff property for $\Delta(e) \geq 0$. Recall that U^P is affine; so let $U^P(\varepsilon) = \alpha\varepsilon - \gamma$. Therefore $E\left[U^P(\varepsilon)|e = e^P\right] = \alpha E\left[\varepsilon|e = e^P\right] - \gamma = \alpha e^P - \gamma \equiv 0$. Moreover,

$$\Delta(e) = \int_0^{e^*} (\alpha\varepsilon - \gamma) dH(\varepsilon|e) = \alpha H(e^*|e)\left[\int_0^{e^*} \frac{\varepsilon dH(\varepsilon|e)}{H(e^*|e)} - \frac{\gamma}{\alpha}\right]$$
$$= \alpha H(e^*|e)\left[e^* - \frac{\gamma}{\alpha} - \int_0^{e^*} \frac{H(\varepsilon|e)}{H(e^*|e)} d\varepsilon\right].$$

Finally, the MLRP implies that $H(\varepsilon|e)/H(e^*|e)$ is decreasing in e for $\varepsilon < e^*$. ■

Supplement to Section 4: refraining from questioning moral imperatives. To show that even questioning an imperative may be unwise, let us return to the basic framework. Assume, for simplicity only, that U^P does not depend on the agent's type (e.g., $U^P(e) = e - \kappa$, where κ is a constant), and thus neither does e^P, defined by $U^P(e^P) \equiv 0$. Let there now be two varieties of the high type, $v_H = v_1$ and $v_H = v_2$, in proportions $1-\lambda$ and λ, so, with average $v \equiv \lambda v_2 + (1-\lambda)\, v_1$, and such that the better type $v_2 > v_1$ is so highly prosocial that, regardless of reputational incentives, he always chooses $a = 1$ when the principal so desires:

$$v_2 e^P - c \geq 0.$$

In contrast, type v_1 will be called "morally fragile". Suppose further that, if the principal issues an imperative instead of disclosing e, the agent can still learn e at an infinitesimal cost, and that this "questioning" of the imperative is observable. We look for an equilibrium in which:

(i) The principal issues an imperative if and only if $e \geq e^P$, as before.

(ii) The high types v_H (whether v_1 or v_2) do not attempt to learn e and conform to the imperative ($a = 1$) when it is issued, while the low type also does not attempt to learn e but picks $a = 0$.

(iii) Were the agent to learn e, an off-the-equilibrium-path event, society would form posterior beliefs $\hat{v} = v_1$.[33]

For type v_1 to obey the imperative in such an equilibriums, it must be that:[34]

$$\int_{e^P}^{1} (c - v_1 e)\, \frac{dF(e)}{1 - F(e^P)} \leq \mu \left[v - \frac{\rho(1-\lambda) v_1}{1 - \rho\lambda} \right]. \tag{A.10}$$

Next, type v_1, if he acquired the information, would reveal his type; he would then pick $a = 1$ if and only if $v_1 e \geq c$. A sufficient condition (a necessary one if the information cost is low enough) for him not to want to acquire the information is

$$\int_{e^P}^{c} (c - v_1 e)\, \frac{dF(e)}{1 - F(e^P)} < \mu\, (v - v_1) = \mu\lambda\, (v_2 - v_1)\,. \tag{A.11}$$

The left-hand side captures the flexibility benefit of being informed, in that the agent does not feel compelled to behave morally when he does not really want to. The right-hand side represents the opprobrium raised by a departure from deontological rule-following, a cost that is borne even if the agent ends up behaving morally: Only morally fragile agents would consider to even question the imperative; neither the highly moral nor the highly immoral (low) types would find any interest in this quest.

Simple computations show that the right-hand side of (A.10) exceeds that of (A.11). Because the left-hand side (A.11) exceeds that of (A.10), condition (A.10) is verified if (A.11) is. Hence:

Provided that (A.11) is satisfied, then there exists an equilibrium in which the principal issues an imperative, and the morally fragile type does not question it, even if the cost of doing so is zero. The morally fragile type mimics the Kantian behavior of the highly moral type by fear of being perceived as a "calculating individual." This behavior is more likely, the more congruent the principal and the higher the ratio of highly moral to morally fragile types.

[33] This follow for instance, from the D1 refinement. Intuitively, type v_1 gains most from the information.

[34] Were he to pool with the v_L type instead, the posterior following $a = 0$ would be $[\rho(1-\lambda)/(1-\rho\lambda)]\, v_1$.

References

Acemoglu, D. and M. Jackson (2015) "History, Expectations, and Leadership in the Evolution of Social Norms, *Review of Economic Studies,* 82(2): 423–456.

Achtziger, A., Alós-Ferrer, C., and A. K. Wagner (2015) "Money, Depletion, and Prosociality in the Dictator Game," *Journal of Neuroscience, Psychology, and Economics*, 8(1): 1–14.

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

Alexander, L., and M. Moore (2015) "Deontological Ethics," *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), Edward N. Zalta (ed.).

Algan, Y., Benkler, Y., Fuster-Morell, M. and J. Hergueux (2016) "Cooperation In A Peer Production Economy: Experimental Evidence From Wikipedia," Sciences Po Working Paper, November.

Alger, I., and J. Weibull (2013) "Homo Moralis–Preference Evolution under Incomplete Information and Assortative Matching," *Econometrica*, 8(6): 2269–2302.

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

Ambuehl, S. (2016) "An Offer You Can't Refuse? Incentives Change What We Believe," University of Toronto mimeo, October.

Andreoni, J. (1989) "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *Journal of Political Economy,* 97: 1447–1458.

_ _ _ _ _ (1990) "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *Economic Journal,* 100: 464–477.

Aquino, K. and Reed II, A. (2002) "The Self-Importance of Moral Identity," *Journal of Personality and Social Psychology*, 83(6): 1423–1440.

Ariely, D., Bracha, A., and S. Meier (2009) "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, 99(1): 544–555.

Ashraf, N., Bandiera, O. and J. Kelsey (2014) "No Margin, No Mission? A Field Experiment on Incentives for Public Services Delivery," *Journal of Public Economics:* 120: 1–17.

Bandura, A. (1999) "Moral Disengagement in the Perpetration of Inhumanities," *Personality and Social Psychology Review*, 3(3): 193–209.

Barrera, O., Guriev S., Henry E. and E. Zhuravskaya "Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics," Sciences Po mimeo, October 2017.

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

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., and J.A. Strongman (1999) “Moral Hypocrisy: Appearing Moral to Oneself Without Being So,” *Journal of Personality and Social Psychology*, 77(3): 525–537.

Battaglini, M., Bénabou, R. and J. Tirole (2005) “Self-Control in Peer Groups,” *Journal of Economic Theory,* 112 (4): 848–887.

Beaman, A. L., Klentz, B., Diener, E., and S. Svanum (1979) “Self-Awareness and Transgression in Children: Two Field Studies,”*Journal of Personality and Social Psychology*, 37(10): 1835–1846.

Bénabou, R. and J. Tirole (2004) “Willpower and Personal Rules,” *Journal of Political Economy*, 112(4): 848–886.

_ _ _ _ _ (2006a) “Incentives and Prosocial Behavior,” *American Economic Review*, 96(5): 1652–1678.

_ _ _ _ _ (2006b) “Belief in a Just World and Redistributive Politics,” *Quarterly Journal of Economics,* 121(2): 699–746.

_ _ _ _ _ (2011a) “Identity, Morals and Taboos: Beliefs as Assets,” *Quarterly Journal of Economics*, 126(2): 805–855.

_ _ _ _ _ (2011b) “Laws and Norms,” NBER Working Paper 17579, November.

_ _ _ _ _ (2016) “Mindful Economics: The Production, Consumption and Value of Beliefs,” *Journal of Economic Perspectives,* 30(3), Summer: 141–164.

Bénabou, R., Falk, A. and J. Tirole (2018) “Eliciting Moral Preferences,” mimeo, June 2018.

Bentham, J. (1789) *An Introduction to the Principles of Morals.* London: Athlone.

Bisin, A., and T. Verdier (2001) “The Economics of Cultural Transmission and the Dynamics of Preferences,” *Journal of Economic Theory*, 97: 298–319.

Bloch, F., R. Kranton, and G. Demange (2016) “Rumors and Social Network,” Working Paper 33.

Bordalo, P., Coffman, K., Gennaioli, N. and A. Shleifer (2016) “Stereotypes,” *Quarterly Journal of Economics,* 131(4): 1753–1794.

Bradley-Geist, J. C., King, E. B., Skorinko, J., Hebl, M. R., and C. McKenna (2010) “Moral Credentialing by Association: The Importance of Choice and Relationship Closeness,” *Personality and Social Psychology Bulletin*, 36(11): 1564–1575.

Brekke, K. A., S. Kverndokk, and K. Nyborg (2003) “An Economic Model of Moral Motivation,” *Journal of Public Economics,* 87 (9-10): 1967–1983.

Brock, J. M., Lange A., and E. Y. Ozbay (2013), “Dictating the Risk: Experimental Evidence on Giving in Risky Environments, ” *American Economic Review*, 103(1): 415–437.

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

Chater, N. and G. Loewenstein (2016) "The Under-Appreciated Drive for Sense-Making," *Journal of Economic Behavior and Organization,* 126(B): 37–154.

Cummiskey, D. (1996) *Kantian Consequentialism.* New York: Oxford University Press.

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.

Dana, J., Weber, R., and J. Kuang (2007) "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preferences for Fairness," *Economic Theory*, 33: 67–80.

Darley, J. M., and B. Latané (1968) "Bystander Intervention in Emergencies: Diffusion of Responsibility," *Journal of Personality and Social Psychology*, 8 (4, Pt.1): 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.

Diener, E., and M. Wallbom (1976) "Effects of Self-Awareness on Antinormative Behavior," *Journal of Research in Personality*, 10(1): 107–111.

Dillenberger, D. and Sadowski, P. (2012) "Ashamed To Be Selfish, " *Theoretical Economics,* 7(1): 99–124.

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 (2009) *The Psychology of Learning and Motivation*, Burlington, VT: Academic Press vol. 50, p. 307–338.

Dohmen, T., Falk, A., Huffman, D. and U. Sunde (2012) "The Intergenerational Transmission of Risk and Trust Attitudes," *Review of Economic Studies*, 79(2): 645–677.

Effron, D. A., Cameron, J. S., and B. Monin (2009) "Endorsing Obama Licenses Favoring Whites," Journal of Experimental Social Psychology, 45(4): 590–593.

Effron, Monin, B., and D. T. Miller (2012) "The Unhealthy Road Not Taken: Licensing Indulgence by Exaggerating Counterfactual Sins," Journal of Experimental Social Psychology, 49(3): 573–578.

Egas, M. and A. Riedl (2008) "The Economics of Altruistic Punishment and the Maintenance of Cooperation" *Proc. R. Soc. B*, 275: 871–878.

Elias, J., Lacetera, N., and M. Macis (2016) "Efficiency-Morality Trade-Offs In Repugnant Transactions: A Choice Experiment," NBER W.P. 22632, September.

Eliaz, K. and Spiegler, R. (2018) "A Model of Competing Narratives," Tel-Aviv University mimeo, November.

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. (2016) "In Face of Yourself - A Note on Self-Image," mimeo.

Falk, A., Fehr, E. and U. Fischbacher (2008) “Testing Theories of Fairness—Intentions Matter,” *Games and Economic Behavior,* 62(1): 287–303.

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

________ (2017) “Diffusion of Being Pivotal and Immoral Outcomes,” Discussion Paper, University of Bonn.

Foerster, M. and J. van der Weele (2018a) “Denial and Alarmism in Collective Action Problems,” Tinbergen W.P. University of Amsterdam, February.

Foerster, M. and J. van der Weele (2018b) “Persuasion, Justification, and the Communication of Social Impact,” Tinbergen Institute W.P. 2018-067/I, August.

Gächter, S. and B. Herrmann (2009) “Reciprocity, Culture and Human Cooperation: Previous Insights and a New Cross-Cultural Experiment,” *Philosophical Transactions of the Royal Society*, 364: 791–806.

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

Gambino, R. (1973) “Watergate Lingo: A Language of Non-Responsibility,” *Freedom at Issue*, 22(7-9): 15–17.

Gerber A., Green, D. and C. Larimer (2008) “Social Pressure and Voter Turnout: Evidence From a Large- Scale Field Experiment,” *American Political Science Review,* 102(1), 33-48.

Gert, B. and J. Gert (2016) “The Definition of Morality,” *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), Edward N. Zalta (ed.).

Gino, F., Norton, M. and R. Weber (2016) “Motivated Bayesians: Feeling Moral While Acting Egoistically,” *Journal of Economic Perspectives*, 30(3), Summer, 189–212.

Gino, F., Schweitzer, M. E., Mead, N. L., and D. Ariely (2011) “Unable to Resist Temptation: How Self-Control Depletion Promotes Unethical Behavior,” *Organizational Behavior and Human Decision Processes*, 115(2): 191–203.

Glaeser, E. L. (2005) “The Political Economy of Hatred,” *Quarterly Journal of Economics*, 120: 45–86.

Gneezy, U., Keenan, E. A., and A. Gneezy (2014) “Avoiding Overhead Aversion in Charity,” *Science*, 346(6209): 632–635.

Goeree, J. K., Holt, C. A., and S. K. Laury (2002) “Private Costs and Public Benefits: Unraveling the Effects of Altruism and Noisy Behavior,” *Journal of Public Economics*, 83(2): 255–276.

Golman, R., Loewenstein, G., Moene, K. and L. Zarri (2016), “The Preference for Belief Consonance,” *Journal of Economic Perspectives*, 30(3), Summer: 165–188.

Gottfredson, M. R. and T. Hirschi (1990) *A General Theory of Crime.* Stanford, Stanford University Press.

Grossman, Z. and J. van der Weele (2017) "Self–Image and Willful Ignorance in Social Decisions," *Journal of the European Economic Association,* 15(1): 173–217.

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

Haidt, J. (2001) "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review*, 108(4): 814–834.

______ (2007) "The New Synthesis in Moral Psychology, " *Science*, 316: 998–1002.

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

Hamman, J., Loewenstein, G., and R. Weber (2010) "Self-interest through Delegation: An Additional Rationale for the Principal-Agent Relationship," *American Economic Review*, 100(4): 1826–1846.

Hare, R. M. (1993) "Could Kant Have Been a Utilitarian?," in Dancy, R. M. *Kant and Critique: New Essays in Honor of W.H. Werkmeister.* Dordrecht, Springer Science & Business Media.

Johnson, R. (2014) "Kant's Moral Philosophy," *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.).

Jordan, J., Mullen, E., and J. K. Murnighan (2011) "Striving for the Moral Self: The Effects of Recalling Past Moral Actions on Future Moral Behavior," *Personality and Social Psychology Bulletin*, 37(5): 701–713.

Juille, T. and D. Jullien (2016) "Narrativity from the Perspectives of Economics and Philosophy: Davis, Ross, Multiple-Selves Models... and Behavioral Economics," GREEG Working Papers Series, No. 2016–19.

Kagel, J. H. and A. E. Roth (1995) Chapter 2 in *The Handbook of Experimental Economics.* Princeton, NJ: Princeton University Press.

Kant, I. (1785) *Grundlegung zur Metaphysik der Sitten.*

_____ (1797) "On a Supposed Right to Lie from Philanthropy," In Kant, I. and M. J. Gregor (1996) *Practical Philosophy*, Cambridge: Cambridge University Press.

Karlsson N., Loewenstein G. and J. McCafferty (2004) "The Economics of Meaning," Nordic Journal of Political Economy, 61–75.

Keeton, R. M. (2015) "'The Race of Pale Men Should Increase and Multiply': Religious Narratives and Indian Removal," in Presser, L. and S. Sandberg (2015) *Narrative Criminology: Understanding Stories of Crime*, New York and London: New York University Press.

Khan, U. and R. Dhar (2006) "Licensing Effect in Consumer Choice," *Journal of Marketing Research*, 43(2): 259–266.

Knoch, D., Gianotti, L. R., Pascual-Leone, A., Treyer, V., Regard, M., Hohmann, M., and P. Brugger (2006) "Disruption of Right Prefrontal Cortex by Low-Frequency Repetitive Transcranial Magnetic Stimulation Induces Risk-Taking Behavior," *Journal of Neuroscience*, 26(24): 6469–6472.

Kranz (2010) "Moral Norms in a Partly Compliant Society," *Games and Economic Behavior,* 68(1).

Lacetera, N., Macis, M. and R. Slonim (2012) "Will There Be Blood? Incentives and Displacement Effects in Pro-social Behavior," *American Economic Journal: Economic Policy,* 4(1): 186–223.

Lagrange, T. and Silverman, R. (1999) "Low Self-Control And Opportunity: Testing The General Theory Of Crime as an Explanation For Gender Differences In Delinquency," *Criminology,* 37: 41-72.

Lazear, E. P., Malmendier, U., and R. Weber (2012) "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied Economics*, 4(1): 136–163.

Levi, P. (1988) *The Drowned and the Saved.* New York: Summit.

Martinsson, P., Myrseth, K. O. R., and C. Wollbrant (2012) "Reconciling Pro-Social vs. Selfish Behavior: On the Role of Self-Control," *Judgment and Decision Making*, 7(3): 304–315.

Mazar N., Amir O. and D. Ariely (2008) "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance,"*Journal of Marketing Research*, XLV: 633–644.

Mazar N. and C.-B. Zhong (2010) "Do Green Products Make Us Better People?," *Psychological Science*, 21(4): 494–498.

McAdams, D. P. (1985) *Power, Intimacy, and the Life Story*, Homewood, IL: Dorsey.

______ (2006) "The Role of Narrative in Personality Psychology Today,"*Narrative Inquiry,* 16(1): 11–18.

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E., and D. Ariely (2009) "Too Tired to Tell the Truth: Self-Control Resource Depletion and Dishonesty," *Journal of Experimental Social Psychology*, 45(3): 594–597.

Merritt, A. C., Effron, D. A., and B. Monin (2010) "Moral Self-Licensing: When Being Good Frees Us to Be Bad," *Social and Personality Psychology Compass*, 4(5): 344–357.

Michalopoulos, S. and M. Xue (2018) "Folklore," Brown University mimeo.

Mill, J. S. (2002) *Utilitarianism: edited with an introduction by Roger Crisp*, New York: Oxford University Press, Originally published in 1861.

Monin, B. and A. H. Jordan (2009) "The Dynamic Moral Self: A Social Psychological Perspective" In Narvaez, D. and D. Lapsley (eds.) (2009) *Personality, Identity, and Character: Explorations in Moral Psychology.* New York: Cambridge University Press.

Monin, B. and D. T. Miller (2001) "Moral Credentials and the Expression of Prejudice," *Journal of Personality and Social Psychology*, 81(1): 33–43.

Mukand, S. and D. Rodrik (2016) "Ideas Versus Interests: A Unified Political Economy Framework," Harvard Kennedy School mimeo.

Mullainathan, S., Shleifer, A. and J. Schwartzstein (2008) "Coarse Thinking and Persuasion," *Quarterly Journal of Economics*, 123(2): 577–619.

Nikiforakis, N. and H.-T. Normann (2008) "A Comparative Statics Analysis of Punishment in Public-Good Experiments," *Experimental Economics*, 11: 358–369.

Oberholzer-Gee, F. and R. Eichenberger (2008) "Fairness in Extended Dictator Game Experiments," The BE Journal Of Economic Analysis & Policy, 8(1): Article 16.

Osgood, J. M. and M. Muraven (2015) "Self-Control Depletion Does not Diminish Attitudes about Being Prosocial but Does Diminish Prosocial Behaviors," *Basic and Applied Social Psychology*, 37(1): 68–80.

Rand, D., Greene, J. D., and M. A. Nowak (2012) "Spontaneous Giving and Calculated Greed," *Nature*, 489(7416): 427–430.

Roemer, J. (2010) "Kantian Equilibrium," *Scandinavian Journal of Economics,* 112(1): 1–24,

Sachdeva, S., I., R., and D. L. Medin (2009) "Sinning Saints and Saintly Sinners: The Paradox of Moral Self-Regulation," *Psychological Science*, 20(4): 523–528.

Shiller, R. (2017) "Narrative Economics," *American Economic Review,* 107, 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.

Sunstein, C. R. (2005) "Moral Heuristics," *Behavioral and Brain Sciences*, 28: 531–573.

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

Tabellini, G. (2008) "The Scope of Cooperation: Values and Incentives," *Quarterly Journal of Economics*, 123(3): 905–950.

Tversky, A. and D. Kahneman (1974) "Judgment under Uncertainty: Heuristics and Biases." *Science,* New Series, 185(4157): 1124–31

Vallacher, R. and M. Solodky (1979) "Objective Self-Awareness, Standards of Evaluation, and Moral Behavior," *Journal of Experimental Social Psychology*, 15(3): 254–262.

Vitz, P. C. (1990) "The Use of Stories in Moral Development," *American Psychologist*, 45: 709–720.

Yanagizawa-Drott, D. (2014) "Propaganda and Conflict: Evidence from the Rwandan Genocide," *Quarterly Journal of Economics*, 129(4): 1947–1994.

Zimbardo, P. (2007) *The Lucifer Effect: Understanding How Good People Turn Evil.* New York, Random House.

For Online Publication Only

Supplementary Appendix: Moral Standards and Admissible Excuses

I. To act or not to act: searching for reasons

We now consider the case where the signal about 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 (*reputational 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. How strong must an excuse be in order to be socially acceptable? And how much stigma is incurred by someone who behaves selfishly without one?

Absent influence motives, we can "zoom in" on a single actor-audience dyad to analyze this issue of equilibrium *moral standards.* The image-enhancement incentive is most obvious in the case of social esteem, but also arises from self-image concerns. Indeed, considerable evidence on motivated cognition documents a tendency for people to process and interpret information in a self-serving fashion.[35] The search for absolving narratives can thus also be interpreted as a form of motivated moral reasoning (Ditto et al. 2009).

Main intuitions. Is producing an excuse for not contributing a good or bad sign about a person's morality? Moral types are highly concerned (v_H) about doing "the right thing," so their search intensity will reflect the *option value(s)* of finding out whether e might be especially high or low –that is, the extent to which the prior distribution is concentrated in the *upper or lower tails.* 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 will thus also factor into their search decisions, but less so than for immoral types (v_L), who are *solely interested* in finding excuses for behaving selfishly. The moral "meaning of excuses" will thus hinge on the balance between the *tail risks* of incorrect decisions and *visibility concerns.*

Formally, suppose that, prior to acting but after learning his type, the agent can obtain a signal $\sigma = e \sim F(e)$ with any probability x, at cost $\psi(x)$, where F is taken here to have full support on $[0, 1]$; 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 cases of indifference. Finally, for any distribution $F(e)$, we define the two conditional moments

$$\mathcal{M}^-(e) \equiv E_F\left[\tilde{e} \mid \tilde{e} \le e\right] \quad \text{and} \quad \mathcal{M}^+(e) \equiv E_F\left[\tilde{e} \mid \tilde{e} > e\right], \tag{B.1}$$

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.

[35] See the articles in the *Journal of Economic Literature's* Symposium on Motivated Beliefs: Bénabou and Tirole (2016), Gino et al. (2016) and Golman et al (2016).

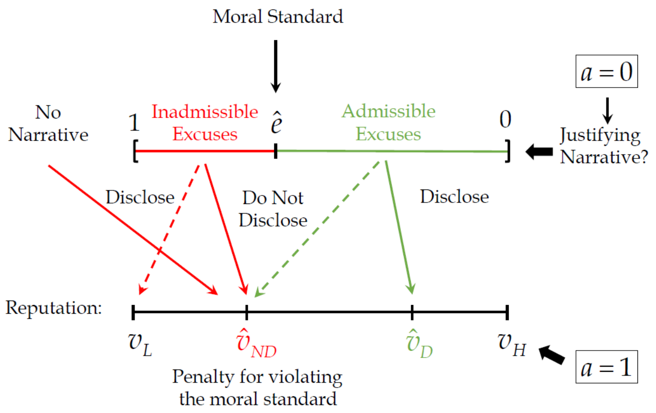

Figure 3: Moral Standards and Narratives. Straight arrows describe equilibrium play, dashed ones off-path deviations

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).

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

$$\hat{v}_D = \frac{\rho x_H v_H + (1-\rho) x_L v_L}{\rho x_H + (1-\rho) x_L}, \tag{B.2}$$

which is independent of e.[36] 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:

$$v_H \hat{e} - c + \mu(v_H - \hat{v}_D) \equiv 0. \tag{B.3}$$

Note that $\hat{e} > e^*$ when $\hat{v}_D > \bar{v}$, or equivalently $x_L < x_H$, and vice versa. We shall denote as $\hat{v}_{ND}$ the audience's posterior when it observes $a = 0$ without a justifying argument. Its value will depend in particular whether the high type's "default" action –his behavior absent any information– is $a = 1$ or $a = 0$, but it must always be that $\hat{v}_{ND} < \hat{v}_D$.[37]

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 *"acceptable"* (e must

[36] 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), $\hat{v}_D$ depends only 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.

[37] 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.

be below $\hat{e}$), and how much *stigma* is incurred for failing to produce one when behaving selfishly ($\bar{v} - \hat{v}_{ND}$). Note from (B.3) 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 much 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, moreover, 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.

II. Looking for "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):

$$v_H(e_0 - e^*) \geq \mu(v_L - \bar{v}) = -\mu\rho(v_H - v_L). \tag{B.4}$$

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 (B.3), while 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 $v_H e - c + \mu v_H$.[38] 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

[38] We assume that the search for reasons and their disclosure are done "on the spot" when confronted with a moral tradeoff (roughly contemporarily with the action choice), whereas the intrinsic and reputational consequences are much longer-lived and thus subject to hyperbolic discounting. If we instead assumed that the value of information is evaluated from the point of view of the ex-ante self, the key formulas and insights would be very similar, except that a term proportional to $c(1/\beta - 1)$ would be subtracted from the right-hand sides of (B.5), (B.8) and (B.11) below. This additional effect naturally makes the high type more averse to information when his default action is $a_L(\emptyset) = 1$, as learning a relatively low e could worsen the temptation to act opportunistically; conversely, it makes him more information-seeking when $a_L(\emptyset) = 0$, as news may provide the missing motivation. In Proposition 8 this makes equilibria more likely to be of the type where $x_H > x_L$ than the reverse, and in Proposition 10, where only the first case is possible, it helps sustain the existence of such an equilibrium.

$$\psi'(x_H) = F(\hat{e})\left[c - \mu(v_H - \hat{v}_D) - v_H\mathcal{M}^-(\hat{e})\right] = F(\hat{e})v_H\left[\hat{e} - \mathcal{M}^-(\hat{e})\right]. \tag{B.5}$$

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

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

leading to

$$\psi'(x_L) = \mu F(\hat{e})(\hat{v}_D - v_L). \tag{B.6}$$

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 (B.2)-(B.6), together with a prior e_0 high enough that (B.4) 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

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

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 (B.5) 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 and associated option values will play an important role here and elsewhere, so we define:

Definition 1. *Given a cutoff $\hat{e} \in (0,1)$, a distribution F_1 is more $\hat{e}$-bottom heavy than another distribution F_2 if $\mathcal{M}^-_{F_1}(\hat{e}) < \mathcal{M}^-_{F_2}(\hat{e})$. Conversely, F_1 is more $\hat{e}$-top heavy than F_2 if $\mathcal{M}^+_{F_1}(\hat{e}) > \mathcal{M}^+_{F_2}(\hat{e})$. If F_1 and F_2 have the same mean and $F_1(\hat{e}) = F_2(\hat{e})$, 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 (e.g., to differ by a mean-preserving spread), while the second precludes it.

Lemma 1.

1. *If F_1 is second-order stochastically dominated by F_2 and $F_1(\hat{e}) \leq F_2(\hat{e})$ (so that $\hat{e}$ is to the right of the intersection point), then F_1 is both more $\hat{e}$-bottom heavy and $\hat{e}$-top heavy than F_2.*

2. *If the likelihood ratio f_2/f_1, or more generally, F_2/F_1, is increasing (resp. decreasing), F_1 is more $\hat{e}$-bottom heavy (resp., $\hat{e}$-bottom heavy) than F_2 at all $\hat{e}$.*

When no confusion results we shall omit the reference to the cutoff, and simply write "bottom (or top) heavy." Formalizing the previous intuitions about each type's incentive to search for excuses, we can now state the following results.

Proposition 8 (prosocial norm). *For any e_0 high enough that (B.4) holds, there exists an equilibrium where moral behavior is the default (uninformed) choice of the high type, 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*

$$e^* - \mathcal{M}^-(e^*) > \mu\rho(v_H - v_L)/v_H, \tag{B.8}$$

   the high type is more likely to search for narratives: $x_H > x_L$, and correspondingly producing one 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 lowers the moral standard ($\hat{e} > e^*$).

2. *If $F(e)$ is sufficiently bottom-light that (B.8) is reversed, it is the low type who is more likely to search for narratives: $x_H < x_L$, and correspondingly producing one 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 raises the moral standard ($\hat{e} < e^*$).*

Intuitively, $e^* - \mathcal{M}^-(e^*)$ scales the option value (relevant only for the high type) 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 (B.8), 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 9. *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 (B.8) 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 high ($\hat{e} < e^*$).*

Intuitively, a higher μ, v_H or $-v_L$ reduces the high type's full-information threshold e^* (by (B.3)), and thus also $e^* - \mathcal{M}^-(e^*)$, since $f/([1-F)$ increasing implies that $0 < d\mathcal{M}^-(e^*)/de^* < 1$; see An 1998). The (normalized) reputational value of excuses $\mu\rho(v_H - v_L)/v_H$, on the other hand, increases in the same way for both types. The net impact of the instrumental and reputational effects thus makes $x_H > x_L$ harder to sustain, and $x_H < x_L$ easier.

III. Looking for "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. Ex-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 (B.2) and the same threshold $\hat{e}$ as in (B.3). Beliefs following $(a = 0, ND)$, however, are now

$$\hat{v}_{ND} = \frac{\rho\,(1 - x_H)\,v_H + (1-\rho)[1 - x_L F(\hat{e})]v_L}{\rho\,(1 - x_H) + (1-\rho)[1 - x_L F(\hat{e})]} > v_L. \tag{B.9}$$

An immoral action without an accompanying excuse is thus less damaging to reputation than in the previous case, since it may now come from an uniformed high type. When there is an excuse, 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

$$v_H(e_0 - e^*) \leq \mu(\hat{v}_{ND} - \bar{v}). \tag{B.10}$$

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, (B.10) 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):

$$\psi'(x_H) = \mu\left(\hat{v}_D - \hat{v}_{ND}\right) + [1 - F(\hat{e})]\left[\mathcal{M}^+(\hat{e}) - \hat{e}\right]v_H, \tag{B.11}$$

$$\psi'(x_L) = \mu F(\hat{e})\left(\hat{v}_D - \hat{v}_{ND}\right). \tag{B.12}$$

so it must always be that $x_H > x_L$, which as noted earlier implies that $\hat{v}_D > \bar{v}$ and $\hat{e} > e^*$.[39]

3. Equilibrium. This is now a quadruplet $(x_H, x_L, \hat{e}, \hat{v}_D) \in [0,1]^3 \times [v_L, v_H]$ satisfying equations (B.2) and (B.11)-(B.12), together with a prior e_0 low enough for (B.10) 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 (which will switch his decision to $a_H = 1$), it is now he who searches more intensively for narratives, compared to the low type.

Proposition 10 (**selfish norm**). *For any e_0 low enough, there exists an equilibrium where abstaining is the default (uninformed) choice of the high type (in particular, (B.10) holds) 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 lowers the moral standard ($\hat{e} > e^*$).*

Interestingly, equations (B.4) and (B.10) can be shown to be compatible over a range of priors, so that both types of equilibria can coexist.

Proposition 11 (**multiple norms and meanings of excuses**). *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 high-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 ($\bar{v} > \hat{v}_D > \hat{v}_{ND} = v_L$).

[39] Clealry, $x_H \geq x_L$. 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 $\mathcal{M}^+(\hat{e}) > \hat{e}$; (B.11) would then lead to $x_H > x_L$, a contradiction.

(ii) A low-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: $v_H > \hat{v}_D > \bar{v} > \hat{v}_{ND} > v_L$).

Summarizing the results in this section, we showed that the key factors determining whether a prosocial or "antisocial" culture tends to prevail are:

(1) Quite naturally, people's prior mean e_0 about whether individual actions have important or minor externalities.

(2) 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" of benevolence –not providing complementary efforts, or even hostile, treacherous, etc. That fear will justify "looking into it," and even when such scrutiny reveals only far less serious concerns (e.g., isolated cases or anecdotes, lowering e only slightly from e^*), such narratives can become socially acceptable reasons for treating 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.

(3) When multiple norms can coexist, the extent to which people want to and/or can coordinate on one or the other. 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 (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), then more prosocial equilibrium $a_H(\emptyset) = 1$ will tend to be collectively preferred, especially if F is top-heavy (and c not too large).

IV- Proofs

Refinements and Uniqueness under Pure Reputation Concerns.

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:*

$$\hat{L}_D = \frac{\rho}{1-\rho}\frac{x_H}{x_L}. \tag{B.13}$$

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

$$\hat{L}_{ND} = \frac{\rho}{1-\rho}\frac{1 - x_H + x_H P(N)}{1 - x_L + x_L\left[P(N) + P(A_1)\right]}. \tag{B.14}$$

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 (B.14) 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 (B.13), that $\hat{L}_D = Q(+\infty)$, which concludes the proof. [40]$\|$

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-equlibrium 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

[40] 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.

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 12. *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 (B.3), 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. ■

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

$$\begin{aligned} E_{\hat{F}_2} - E_{\hat{F}_1} &= \int_0^{\hat{e}} [\hat{F}_1(e) - \hat{F}_2(e)]dx + \left[e(\hat{F}_2(e) - \hat{F}_1(e))\right] |_0^{\hat{e}} \\ &= \int_0^{\hat{e}} \left[\frac{F_1(e)}{F_1(\hat{e})} - \frac{F_2(e)}{F_2(\hat{e})}\right] dx \geq \frac{1}{F_1(\hat{e})} \int_0^{\hat{e}} [F_1(e) - F_2(e)]dx \geq 0, \end{aligned}$$

where the last inequality follows from $F_1 \preccurlyeq_{SOSD} F_2$. Thus, $\mathcal{M}^-_{F_2}(\hat{e}) = E_{F_2^*} \leq E_{F_1^*} = \mathcal{M}^-_{F_1}(\hat{e})$. Similarly, we have

$$\mathcal{M}^+_{F_1}(\hat{e}) - \mathcal{M}^+_{F_2}(\hat{e}) = \int_{\hat{e}}^1 \left[\frac{F_2(e)}{1 - F_2(\hat{e})} - \frac{F_1(e)}{1 - F_1(\hat{e})}\right] dx \geq \frac{1}{1 - F_2(\hat{e})} \int_{\hat{e}}^1 [F_2(e) - F_1(e)] \, dx \geq 0.$$

(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 $\hat{e} \in [0, 1]$, integration by parts yields:

$$\mathcal{M}^-_{F_1}(\hat{e}) - \hat{e} = E[X - \hat{e}|X \leq \hat{e}] = \int_0^{\hat{e}} z\hat{F}_1(z)dz - \hat{e} = -\int_0^{\hat{e}} \frac{F(z)}{F(\hat{e})} dz = -\left(\frac{\partial}{\partial \hat{e}}\left[\ln \int_0^{\hat{e}} F(z)dz\right]\right)^{-1}.$$

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

Proof of Proposition 8 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 (B.3), while (B.5)-(B.6) imply that $x_L \leq x_H$ if an only if:

$$\mathcal{M}^-(\hat{e})v_H \leq c - \mu(v_H - v_L). \tag{B.15}$$

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.[41] Rewriting the inequality slightly using (3) yields the reverse of (B.8).

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 (B.15) is reversed:

$$\mathcal{M}^-(\hat{e})v_H \geq c - \mu(v_H - v_L), \tag{B.16}$$

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 (B.8).

It only remains to prove that an equilibrium with $a_H(\emptyset) = 1$ exists whenever (B.4) is satisfied. Equation (B.3) 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 (B.5)-(B.6) 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, (B.2) 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 10 Each type's expected utility from a search intensity x are now

[41] 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).

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

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 (B.3) 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 (B.11)-(B.12) 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, (B.9) 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 (B.10) must then hold, so all equilibrium conditions are satisfied. ■

Proof of Proposition 11 Let $\underline{e}_0$ be the value of e_0 that makes (B.4) an equality; for all $e \geq e_0$, there exists an equilibrium with $a_H(\emptyset) = 1$. Turning to conditions for an equilibrium, let $\hat{v}_{ND}(e_0) \in [v_L, \bar{v})$ denote any fixed point of the mapping defined by equations (B.3), (B.11)-(B.12) and (B.9); we saw in the proof of Proposition 10 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 (B.10). 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 (B.9), it suffices that $x_H(e_0)$ remain bounded away from 1, and from (B.11) this is ensured as long as $\psi'(1) = +\infty$, since the right-hand side of (B.11) is bounded above by $\mu(v_H - v_L) + v_H[\mathcal{M}^+(\hat{e}) - \hat{e}] < \mu(v_H - v_L) + v_H$. ■

References

An, M. (1998) "Logconcavity versus Logconvexity: A Complete Characterization," *Journal of Economic Theory*, 80, 350-369.