Discussion Paper Series – CRC TR 224

Discussion Paper No. 070
Project A 01

Narratives, Imperatives, and Moral Reasoning

Roland Bénabou*
Armin Falk**
Jean Tirole***

February 2019

*Department of Economics and Woodrow Wilson School of Public and International Affairs, Princeton University. 286 Julis Romo Rabinowitz - Princeton, NJ 08544, USA, e-mail: rbenabou@princeton.edu
**Economics Department, University of Bonn, Adenauerallee 24-42, 53113 Bonn, Germany, e-mail: armin.falk@briq-institute.org
***Toulouse School of Economics, Manufacture de Tabacs, 21 Allée de Brienne, 31015 Toulouse Cedex 6, France, e-mail : jean.tirole@tse-fr.eu

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Narratives, Imperatives, and Moral Reasoning

Roland Bénabou, Armin Falk, and Jean Tirole

July 2018

Abstract : By downplaying externalities, magnifying the cost of moral behavior, or suggesting not being pivotal, exculpatory narratives can allow individuals to maintain a positive image when in fact acting in a morally questionable way. Conversely, responsibilizing narratives can help sustain better social norms. We investigate when narratives emerge from a principal or the actor himself, how they are interpreted and transmitted by others, and when they spread virally. We then turn to how narratives compete with imperatives (general moral rules or precepts) as alternative modes of communication to persuade agents to behave in desirable ways.

1 Introduction

1.1 Moral decisions and moral reasoning

What is the moral thing to do? The aim of this paper is of course not to answer that immemorial 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" pervade the social and political discourse, often trumping any argument of economic efficiency (bans on "immoral" transactions, trade wars, undeservedness of some group, etc.). And, of course, everyone experiences inner struggles over these issues.

Moral arguments can provide reasons for what one "should do," or on the contrary justifications for acting according to self-interest, under given circumstances. Alternatively, they may be broad "fiat" prescriptions, dictating a fixed behavior across most situations, without explaining why. We refer to these two classes as moral *narratives* and *imperatives,* respectively, and explore how they combine with more standard motivations such as social preferences, self-control, and image or identity concerns to shape behaviors and ultimately favor the emergence of pro-or anti-social norms. Concerning narratives, we highlight two main questions. The first one is that of *viral transmission:* what types of social structures lead exculpatory versus responsibilizing rationales to spread widely, or remain clustered within subgroups? The second question is that of *moral standards:* how tolerant is a society of excuses for self-interested behavior (how compelling do they have to be), how much stigma is borne by those who fail to produce one, and how hard do people search for receivable arguments? Turning next to imperatives, the first key issue is how they work: what characteristics confer someone the moral *legitimacy* to issue them and have (certain) others 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 sometimes influence its course; they are "instrument[s] of mind in the construction of reality" (Bruner 1991, p. 6). Narratives are viewed as central features of all societies by many disciplines including anthropology, psychology, sociology, history and the humanities; they are now starting to attract attention by economists.[1]

Given such a broad concept, it is useful to first distinguish two main types, and roles, of narratives. The 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, Chaterand and Loewenstein 2010). This drive reflects a strong need for predictability, serving both planning and anxiety-reduction purposes, and probably involves evolutionarily selected programs for pattern seeking. The second sense, which we explore here, is that of *narratives as rationales or justifications.* These may be arguments shaping standard

[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. For recent discussions of narratives in economics, see Shiller (2017) and Juille and Jullien (2016).

economic decisions, e.g., adverting slogans like "Because you are worth it," or the recurrent narrative identified by Shiller (2017) in many real-estate bubbles that, for sure, "They are not making any more land." The most important narratives, however, pertain to actions with *moral or social implications,* namely those involving externalities/internalities and (self-) reputational concerns. It is on such rationales for what one "should do" (or not) that we shall focus. Accordingly, we define a moral narrative as any news, story, life experience or heuristic that has the potential to alter an agent's beliefs about the tradeoff between private benefits and social costs (or the reverse) faced by a decision-maker, who could be himself, someone he observes, or someone he seeks to influence. It may be received fortuitously, searched for and thought of by the individual himself, or strategically communicated by someone else.

Having the potential to alter beliefs does not necessarily require the story to have any truth or relevant informative content, nor that receivers respond to it with fully rational (Bayesian) updating. Of course such may be the case, as with (say) hard evidence on second-hand smoking or prevalent corruption. All that matters, however, is that there be a *perceived* "grain of truth" prompting persuasion; indeed, some of the most successful narratives, involving negative ethnic and gender stereotypes, are plainly wrong. Alternatively, the fact itself may be correct but provide a very incomplete picture and therefore be potentially misleading if one jumps to the conclusion: "This year's frigid winter proves that global warming is a hoax," "I have a friend who...", etc. Vivid life experiences, simple and 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. Base-rate neglect, confusion between correlation and causality, and motivated reasoning also offer many avenues for narratives to "work" where, under full rationality, they should not.

We see *imperatives* as located at the opposite end of the independent-persuasiveness spectrum: whereas narratives either are, or at least act like, hard information, imperatives are entirely soft messages of the type "thou shalt (not) do this," seeking to constrain behavior without offering any reasons why (other than a tautological "that is just wrong," or "because I say so.") Thus, while a narrative can by itself alter an individual's beliefs and actions, independently of where it came from, imperatives are relationship-dependent: whether such rules are obeyed, ineffective, or backfire depends on whether their author is regarded as trustworthy and benevolent, neutral, or adversary. And while 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 influence is of course a simplifying first step. In practice, most moral arguments lie along a continuum between these polar opposites, or explicitly combine the two.

1.3 Formalization and main results

Our starting point is a simple, workhorse model of individual moral decisions. Following the utilitarian philosophical 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. Individuals differ in their intrinsic valuations for providing the externality and, in line with the considerable evidence that people strive to maintain a positive self-concept and social image (e.g., Aquino and Reed II 2002;

Mazar et al. 2008; Monin and Jordan 2009), they derive reputational benefits from being perceived, or seeing themselves, as having high moral values. Finally, reflecting the fact that opportunistic behavior often arises from momentary temptations, agents may (but need not) have imperfect self-control when trading off the personal costs of their actions against future social and reputational consequences. Quite naturally, an individual is more likely to act morally the higher are the perceived externality, his image concern and willpower, and the lower his initial level of reputation (keeping actual preferences fixed). We discuss how these simple predictions match a wide range of experimental evidence from both psychology and economics.

This basic framework then sets the stage for *narratives*, introduced as arguments or rationales about the moral consequences, i.e., the social costs and benefits, of a person's actions. Eschewing any specific channel, whether "rational" or "behavioral," through which an argument may sway beliefs, we focus on *why and how* people use such moral-persuasion devices, on how observers draw inferences from both speech and action, and on the social norms and social discourse that result. Two main categories of arguments or rationales 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 examples and experimental evidence on both types: common "neutralization" rationales (Sykes and Matza 1957) 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?").

Narratives clearly also involve a critical social-contagion aspect. To analyze this phenomenon, we embed the basic model into a linear network that mixes different types of individuals (men and women, majority and minority, rich and poor, etc.), each of whom may observe the actions of their predecessor, receive a narrative from him or her, and transmit it to their successor. This simple setup first brings to light two key mechanisms determining the spread of different moral arguments through a population, which we term the *reputational* and *social-influence* motives. As before, an actor who learns of a narrative justifying selfish behavior has a reputational 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 also act egoistically and invoke the rationale to his own successor, and so on: the initial disclosure thus triggers a chain of bad behaviors. Conversely, sharing a positive, responsibilizing narrative forces one to act morally, but has the "multiplier" benefit that the successor may not just be convinced of doing the same, but will then also pass on the responsibility argument to his next neighbor, hence a cascade of prosocial behaviors. We study the interplay of these offsetting incentives, show that negative narratives (i.e., disclosures thereof) are *strategic substitutes* while positive ones are *complements,* and analyze how the equilibrium outcome depends on the networks' mixing of individuals who take a type-indicative action and others who don't – say, men's behavior with respect to women in the workplace.

We show in particular how different norms emerge, in which either other-regard or self-indulgence is the default. In the "moral" equilibrium doing the right thing "goes without saying," and conversely abstaining requires an excuse, so negative narratives are the ones that will get passed on (when they occur) and affect behavior. In the "amoral" equilibrium, self-indulgence is the default, but excuses remain reputationally valuable and thus again circulate.

So now will positive narratives, however, propagated by passive and high-morality active agents to induce others to action. Most importantly, we show that, in any type of equilibrium, more *mixed interactions raise prosocial behavior.* Because agents whose morality is not "in question" have no need for excuses, they act both as "firewalls" limiting the diffusion of exonerating narratives and as "relays" for responsibilizing ones. In the latter case this also encourages high-morality actors to do the same, as the positive narrative will travel further. We also show that greater mixing (with proportions fixed) results in beliefs that are both less clustered and *less polarized across the two groups.* Conversely, a high degree of spatial correlation within the network causes very different types of narratives to circulate within each of the two populations – a form of mutual stereotyping. In the gender example, for instance, 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 social setting.

Besides exogenous sources and strategic third parties, narratives can also originate in an individual's own *search for reasons* to behave one way or another. The fact that someone produces an excuse for acting in line with self-interest is then indicative that they may have sought one, or more generally "looked into the question," and this itself changes the view of their motivations –as in: "Isn't it interesting how he always has a good excuse not to help..." This leads us to investigate the question of a group or society's *moral standards:* what justifications it deems acceptable or unacceptable, how it judges people who do or do not have "good enough" excuses, and the *different norms* that can result for both behavior and discourse. The key factors determining whether a prosocial culture, an antisocial one, or both, will emerge are shown to be: (i) people's prior mean about whether individual actions have important or minor externalities; (ii) the *tail risks* (or option values) in the uncertainty surrounding that mean. For instance, even a very small probability that some out-group may "deserve" harsh treatment (e.g., it has only a minor impact on their welfare, and a positive one of that of the ingroup) can justify scrutinizing their merits or demerits, so that even when this reveals only much weaker justifications (e.g., anecdotes), having those on hand becomes socially acceptable. There are now "excuses for having excuses," and as a result moral standards are lower. In the process of the analysis, we also provide news results on ranking probability distributions according to upper or lower conditional moments (termed "bottom- or top-heaviness").

To study imperatives and how they can emerge even in a utilitarian framework, we next enrich the earlier "influence" channel, but focusing now on communication between a single principal (parent, religious leader, etc.) and an agent. The principal cares for the welfare of society at large and/or that of the agent, and can issue either a narrative or an imperative. As before, narratives are signals, or messages interpreted as such, about the externalities resulting from the agent's choices; they could also be about private costs or visibility of his actions. In contrast, an imperative is a precept to act in a certain way, without looking to consequences for reasons. The analysis reveals several costs and benefits of imperatives, relative to narratives. On the cost side, imperatives are effective only if issued by principals with moral authority (perceived competence and benevolence), whereas narratives can persuade irrespective of their source. This suggests why imperatives are rarely used in the political arena –where narratives instead prevail– but are more common in parent-child interactions or in religious writings, such as the Ten Commandments. Another restrictive feature of imperatives, similar to the rule vs. discretion idea, is that they impose rigidity on the 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 to interpretation uncertainty and to the threat of counter-arguments. When effective, they also expand the range over which the principal induces desired behaviors by the agent, resulting in the provision of greater externalities. This offers a possible explanation for why imperatives typically consist of unconditional prescriptions with a large scope, such as not to lie, steal or kill. Finally, we show that under certain conditions, the use of imperatives rather than narratives is more likely if agents suffer from self-control problems. Like other personal rules (Bénabou and Tirole 2004) or moral heuristics (Sunstein 2005), they help individuals otherwise prone to impulsive decisions achieve their long-term goals, but occasionally sometimes misfire, due to their rigidity.

The paper is organized as follows. The remainder of this section discusses related literature. Section 2 introduces the basic framework and shows how core preferences, self- or social-esteem concerns, and beliefs about the consequences of one's actions, potentially shaped by narratives, jointly determine moral conduct. Section 3 analyzes how the transmission of arguments on a simple network is shaped by the tradeoff between reputation and influence concerns, and how the degree of mixing between different types affects the average behavior and the polarization of beliefs. Section 4 focuses on people's endogenous search for reasons to justify behaving pro- or anti-socially, and the resulting emergence of different moral standards and norms. The legitimacy of imperatives and the tradeoffs between using them versus narratives to influence moral behavior are analyzed in Section 5. Section 6 concludes with directions for further research. All proofs are in appendices.

1.4 Related economics literature

The paper relates to and extends 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 an agent taking some costly action, we add the new one of invoking potential 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 the topics of public goods and learning in networks. Most of this literature features agents who learn mechanically from their neighbors (e.g., DeGroot model), and/or spontaneously emit information toward them. Our paper is thus most connected to the recent strand in which communication is instead strategic (e.g., Hagenbach and Koessler 2010, Galeotti et al. 2013, Ambrus et al. 2013, Bloch et al. 2016).

Because the arguments that individuals produce and circulate here are specifically moral narratives and imperatives, the paper also contributes 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).

From a different perspective, the paper also ties into the literature on the cultural trans-

[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)

mission of values, beliefs and ideas (e.g., Bisin and Verdier 2001, Bénabou and Tirole 2006b, Tabellini 2008, Dohmen et al. 2012). In our case, the focus is on ideas, beliefs and norms about what is morally right or wrong. Finally, the importance of informal "stories" in the formation of economic actors' beliefs has recently been emphasized by Shiller (2017), and a few other recent papers are beginning to explore the roles played by narratives, memes or folklore (e.g., Mukand and Rodrik 2016, Barrera et al. 2017, Michalopoulos and Xue 2018).

Most closely related to ours is independent work by Foerster and van der Weele (2018). Their model has bilateral communication between two agents, each endowed with a prosociality type and an imperfect signal about the value of the externality. Each one sends a message to the other before they both act, and communication is cheap-talk rather than through disclosure of signals. Image concerns then generate an incentive to understate the externalities from behaving selfishly ("denial") , while the desire to induce more pro-social behavior by the other player pushes towards exaggerating them ("alarmism"). This broadly parallels, but in the form of distorted reports, the ways in which the reputational and influence motives in our model lead to the selective disclosure of negative versus positive narratives. Foerster and van der Weele also show that image concerns work against "hypocrisy" (exaggerating signals while behaving selfishly), so that any equilibrium with information must involve some denial, and that the latter may improve welfare. Besides the different information technologies in the two models, we focus on the endogenous search for reasons or arguments and, especially, on issues related to the dynamics of (serial) communication between many agents: strategic complementarity or substitutability, viral transmission, polarization, and the effects of social mixing. We also investigate imperatives, which are unidirectional cheap talk.

2 Basic Model

2.1 Moral decisions and moral types

The basic model 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. We will also allow agents to have imperfect willpower at the moment of choice: the ex-ante cost c of "doing the right thing" is momentarily perceived as c/β, where $\beta \leq 1$ is the individual's degree of self-control or (inverse) hyperbolicity.

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 enjoyment or 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-esteem and 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 forgotten, or only imperfectly accessible in retrospect; only the hard facts of the decision, $a = 0$ or 1, can be reliably recalled. Either way, at the point where he chooses his action a, an agent of type $v = v_H, v_L$ has expected utility

$$\left(ve - \frac{c}{\beta}\right) a + \mu \hat{v}(a), \tag{1}$$

where $\hat{v}(a)$ is the expected type conditional on action $a \in \{0, 1\}$ and μ measures the strength of self or social image concerns, common to all agents. 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.

Assumption 1.

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

The first inequality says that not contributing is a strictly dominant strategy for the low, "immoral" type: he prefers to abstain even when the social and reputational benefits are both maximal, $e = 1$ and $\hat{v}(1) - \hat{v}(0) = v_H - v_L$. The second inequality says that both types pooling at $a = 0$ is not an equilibrium when the externality is maximal ($e = 1$): the high type would then want deviate to $a = 1$, even at minimal image gain $v_H - \bar{v}$. Consequently, in a pooling equilibrium at $a = 0$ 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

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

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 choose the equilibrium that is best for both types, namely the no-contribution pooling equilibrium. Indeed, the separating equilibrium yields lower payoffs: $\mu \cdot v_L < \mu\bar{v}$ for the low type and $v_H e - c/\beta + \mu v_H \leq \mu\bar{v}$ for the high one.[3] 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*

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

Immoral behavior is encouraged by a low perceived social benefit e, a high personal cost c or low

[3] Pareto dominance is understood here as better for both types of a single individual. Since $a = 1$ has positive externalities, this will generally be different from (possibly even opposite to) Pareto efficiency understood in the sense of making everyone in society (a large group) better off. If we instead selected the separating equilibrium there would be more of an alignment, but one would have to deal with more complex, mixed strategies. The comparative statics of interest would remain the same, however.

degree of self control β, and a weak reputational concern μ.[4]

We next discuss how these predictions align with a broad range of empirical evidence, though we also point out contrary findings and issues on which the debate remains unsettled. The purpose is not to formally test this basic model, but to verify that it is empirically sound before proceeding to build further upon it.

2.2 Related evidence

1. Social and self-image concerns (μ) play a central role here, as in a broad class of related signaling models. Increased visibility is thus predicted to induce more moral behaviors, and this is indeed found in a wide variety of contexts, ranging from charitable contributions (e.g., Ariely et al. 2009, Della Vigna et al. 2012, Ashraf et. al. 2012) to public goods provision (Algan et al. 2013), voting (Gerber et al. 2008) and blood donations (Lacetera and Macis 2010). A related literature has provided (mixed) evidence that displaying images of eyes increase prosocial behaviors, relative to showing neutral images, presumably for reputational reasons (see, e.g., Haley and Fessler 2005). Finally, the key role of *attributed intentions* (versus final outcomes) in determining social sanctions and rewards, is well established (e.g., Falk et al. 2008).

Self-image concerns have very similar effects, as raising an agent's awareness of his own choices or/and prevailing ethical standards also corresponds to increases in μ. Experimental evidence indeed documents that more salient moral standards and greater self-awareness (especially in conjunction with each other) lead to greater fairness and honesty (Batson et al. 1999) and less 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). In this experiment, 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 exerting the negative externality is significantly lower, by about 25%.

2. Initial self or social image (ρ or $\bar{v}$). The model predicts a higher likelihood of unethical choices in the presence of a high initial reputation, a set of behaviors that corresponds to what social psychologists term "moral licensing" (for the reputation-rich) and "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).[5]

[4] When $e > e^*$, the separating (or "moral") equilibrium, in which the high type chooses $a = 1$, is the unique one. When $e \leq e^*$, the pooling (or "immoral") equilibrium in which both types choose $a = 0$ exists and is better for both types, and thus selected by our criterion.

[5] In Monin and Miller (2001), subjects who had been given an opportunity to demonstrate non-prejudiced attitudes were subsequently more likely to express the belief that a police job is better suited for a White than a Black person. In Effron et al. (2009), after being given the opportunity to endorse Barack Obama participants were more likely to favor Whites than Blacks for a job. In Sachdeva et al. (2009) participants donate less (more) to a charity after being asked to write a personal story including positive (negative) own traits.

3. Self-control (β). Morally demanding decisions often imply a trade-off between immediate gratification (quick money, letting off steam, etc.), and costs accruing in the future in terms of self-image, social reputation, or outright punishment. Martinsson et al. (2012) find that participants who report generally having low self control make more selfish allocations in dictator games. Achtziger et al. (2015) show that when subjects are experimentally "ego depleted" (a manipulation that consumes self-control resources and favors impulsive behavior; see Baumeister et al. 2007), they again share significantly less money in a dictator game. Related experiments shows that depleted self-control also fosters dishonesty (Gino et al. 2011, Mead et al. 2009), criminal behavior (Gottfredson and Hirschi 1990) 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).[6]

4. Costs (c) That moral or 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 is inversely related to the level of cooperation (Goeree et al. 2002, Gächter and Herrmann 2009). Likewise, the willingness to exert altruistic punishment in public goods games with a subsequent sanctioning stage decreases in the cost of punishment (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 a type of "deontological" behavior that we examine in our companion paper (Bénabou, Falk and Tirole 2018).

5. Social externality (e). That prosocial choices are generally sensitive to the implied consequences is well documented in the literature on cooperation and voluntary contribution to public goods (Kagel and Roth 1995, Chapter 2). In a study that cleanly disentangles higher external return (gain 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 overhead 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 that –for a given cost– yield the highest social return. We model agents' preferences in line with this notion and the above evidence, but also make note of two lines of research emphasizing instances of insensitivity to consequences. One stems from impure altruism or "warm glow," where utility is derived from the act as such, not from consequences (e.g., Andreoni 1989, 1990). The other, which is the focus of our above-mentioned companion paper, is a a stated unwillingness to enter moral trade-offs altogether, often referred to as deontological or Kantian reasoning.

[6] We should also note that there generally remains a debate about whether it is selfishness or prosociality that constitutes the first and instinctive moral impulse (or "prepotent response"), which other mental processes may then control or override. Thus, Rand et al.'s (2012) experiments suggest that people are predisposed to acting prosocially, and become self-interested only after some reflection and reasoning. These findings have been challenged in later replications (e.g., Boowester et al. 2017), and the question remains empirically unsettled.

2.3 Seeking or avoiding moral choices

As seen in Proposition 1, image concerns induce the agent to behave more prosocially. From his perspective this has both costs and benefits. Depending on which ones dominate he may seek out, on the contrary actively shun, situations that put his morality to the test –a type of behavior often called *dilemma avoidance.*

Let us therefore ask when, on average (or, behind the veil of ignorance about one's type), a population would a priori prefer to face a restricted choice $\{a = 0\}$, generating average utility $\mu\bar{v}$, or a moral decision $a \in \{0, 1\}$, with image at stake. In the latter case, ex-ante utility is

$$U \equiv \rho U_H + (1-\rho)U_L = E[(ve - c)a + \mu\hat{v}(a)] = E[(ve - c)a] + \mu\bar{v}, \tag{4}$$

by the martingale property of beliefs. Comparing these ex-ante preferences to the ex-post ones given by (1), two wedges are apparent:

(a) The *self-control problem* creates undesirable impulses to behave immorally –cheating, retaliating against or yelling at friends, relatives, colleagues, etc.– which self and social image concerns help counteract. Investing in reputation-sensitive capital, taking advantage of or even seeking out settings where ones' morality will be under scrutiny (effectively raising μ) are then valuable strategies to counteract such temptations.[7]

(b) *Image concerns* tend to generate wasteful signaling: esteem is a purely "positional good," in fixed total supply $\bar{v}$. Moral-choice opportunities may allow the high type to distinguish himself and look better than average, but the low type will correspondingly look worse – a zero-sum game (given the benchmark assumptions of linear reputational gains and a common μ).

Oversignaling occurs when $e > e^*$ but $\rho(v_H e - c) + \mu\bar{v} < \mu\bar{v}$, or equivalently $v_H e < c$. In this case an agent would want to avoid situations where he might feel compelled to act prosocially.[8] Conversely, there is undersignaling when $e \leq e^*$ but $v_H e > c$.[9]

Proposition 2 (avoiding or seeking the ask). *Ex ante, the agent will:*

1. *"Avoid the ask" out of concern that, ex post, he would otherwise oversignal, when $e^* < e < c/v_H$. This occurs for a nonempty range of externalities when μ or β is high enough:*

$$c\left(\frac{1}{\beta} - 1\right) < \mu(v_H - \bar{v}). \tag{5}$$

[7] On the related use and monitoring of self-enforcing "personal rules," see Bénabou and Tirole (2004).

[8] See also Dillenberger and Sadowski (2012) for an alternative formulation, based on temptation preferences rather than signaling concerns.

[9] 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 $(v_H e - c)/\beta + \mu(v_H - \bar{v}) > 0$. 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.

2. *"Seek out the ask" even at some cost, as a means of commitment, when β is low enough:*

$$\max\{c/v_H, e^*\} < e < c/\beta v_H.$$

3. *Be concerned about undersignaling, and thus seek greater exposure (increasing μ), whenever $c/v_H < e \leq e^*$. This occurs for a nonempty range of externalities whenever (5) is reversed.*

Both seeking and avoiding moral scrutiny are observed in practice, and display patterns broadly in line with the corresponding parameter configurations. For help with self-control problems through increased social monitoring and feedback, people join and rely on religious organizations and discussion groups such as Alcoholics Anonymous (see Battaglini et al. 2005). Conversely, they tend to avoid situations where social pressure would lead them to be excessively generous, relative to the "real" stakes. 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 see a beggar too closely. 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, Loewenstein and Weber 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.

2.4 Exoneration and responsibility: narratives as justifications

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. This is already apparent from Proposition 1, where actor and observer share the same belief, but in more general settings the two may of course differ. This is where narratives and other communicable rationales about what constitutes moral or immoral behavior come into play.

We distinguish two main types of signals or arguments affecting agents' actions.

(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. Since they are detrimental to other parties, we shall also call them *negative narratives*. They operate through many exculpatory or "neutralization" strategies (Sykes and Matza 1957), such as: (a) downplaying the harm**;** (b) blaming the victims; (c) denying agency and responsibility; (d) appealing to higher loyalties like religious values or missions that justify harming 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. Since they lead to the provision of beneficial externalities, they will also be referred to as *positive narratives.* Examples include: (a) appeals to moral and religious precepts, 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 on ones' choices (God, Adam Smith's "impartial spectator within the breast," "Your mother if she could see you," etc.).

As some of the examples show, these stories need not be true in an objective sense to nevertheless powerfully influence a person's behavior and judgment (see, e.g., Haidt et al. 2009). They could be any of: (a) 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.[10] The essential feature for the *positive* analysis, which is our main focus, is 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.[11] For *normative* conclusions, of course, their veracity or falsehood matters importantly.

Formally, a narrative is a signal or message –whether genuine information, frame, cue,

[10] On (b), see Tversky and Kahneman (1974, 1981), Mullainathan et al. (2008) or Bordalo et al. (2016). Examples of (c), include, on the negative side,"easy-fix" solutions (e.g., tax cuts will pay for themselves), wishful-thinking and conspiracy theories. On the positive one, consider the ubiquitous "Imagine everyone did this". Given that one's action will *not* become universal law (and a small marginal pollution "makes no difference" overall), why is this argument (also related to Kant's categorical imperative) so powerful? Our conjecture is that it may be hard for an individual to figure out whether a small externality e (dropping a paper on the ground) justifies a small cost c (carrying it to the next garbage can), implying a moral duty. The narrative magnifies the salience of both (e.g., one envisions dirty cities), facilitating the comparison.

[11] 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 that actual ones, and that correcting the facts does nothing to undo these effects.

rhetorical devices, etc.– that, when received by an agent, will move his expectation of the externality from the prior mean e_0 to some value e, drawn from $[0,1]$ according to a prior distribution $F(e)$. 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. Realizations of e will be "negative narratives" or "excuses" when they reduce moral behavior for some range of priors, but never increase it. Conversely, "positive narratives" or "responsibilities" are those with the potential to increase moral behavior, but not to decrease it. When no narrative is received we denote the signal as $\emptyset$, and the agent just keeps his prior e_0.

Where do narratives come from? The probability with which a narrative is obtained, denoted x, can be exogenous or endogenous. The first case arises for instance when agents receive messages from a "narrative entrepreneur" who seeks to induce a given behavior.[12] Normalizing $x \equiv 1$, Proposition 1 directly applies, with negative narratives corresponding to $e \leq e^*$ and positive ones to $e > e^*$. In the second case, it is useful to distinguish two channels, though in practice they often operate jointly. Section 3 below focuses on the *transmission* channel, in which each agent be may be told of the narrative by a friend, neighbor or colleague, and then decide whether to pass it along to someone else. Section 4 will then turn to the *production* channel, where the signal arises from an agents' own search for reasons to act morally or selfishly, leading again to a decision of whether or not to disclose (or dwell on / or keep in mind) the arguments he came up with.

3 The Viral Transmission of 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. The same two effects operate in reverse for positive narratives: sharing information suggesting that one knowingly chose some that imposes significant social harm is highly detrimental to reputation; conversely, the social-influence effect is now positive, as awareness of consequences promotes others' moral behavior.[13]

[12] For instance, Glaeser (2005) formalizes the notion that politicians seek to expand their political power by creating and broadcasting stories that sow hatred against a minority.

[13] That sharing a negative narrative (low e) is beneficial to one's reputation, and conversely sharing a positive one (high e) is detrimental to it, is a general insight, not limited to the case where the chosen action is the selfish one, $a = 0$. Since intrinsic motivation is ve, acting prosocially 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 since $a = 1$ is sufficient to reveal the high type, but more generally they would.

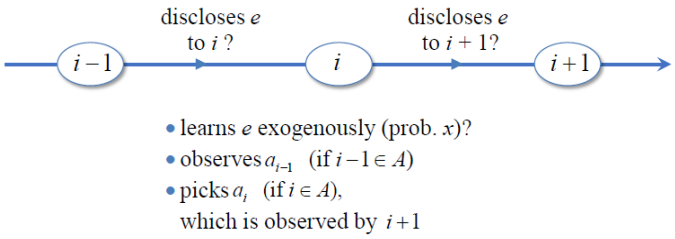

Figure 1: Viral Transmission of Narratives

3.1 Sharing narratives on a simple stochastic network

1. Setup. Let there be a countable set of individuals $i \in \mathbb{N}$, 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, everyone acts, but P agents' behavior is unobservable.

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

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

where A and P respectively denote the sets of active and passive individuals. 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 his predecessor $i-1$ was active or passive. The following "time symmetry" implication of (6), resulting from the fact that the invariant distribution of types is 50-50, will therefore be useful:

$$\Pr[i-1 \in A \mid i \in A] = \Pr[i-1 \in P \mid i \in P] = \lambda. \tag{7}$$

Agents' preferences remain unchanged: among active individuals, a proportion ρ have moral type H and the remaining $1-\rho$ moral type L (for whom $a=0$ is a dominant strategy), and all share the same reputational concern μ with respect to their audience, which for each agent i is simply his successor $i+1$. For simplicity there is, at any given time, a single potential signal or "narrative" about the importance of externalities, received independently by each agent i with constant probability x, and with support F restricted to two values: e equals e_- (probability f_-) or e_+ (probability f_+), with $e_- < e^* < e_+$.

While abstracting from simultaneous contests between opposing rationales, this framework will nonetheless allow us, by averaging across realizations or periods, to examine how the conversations, beliefs and behavior of a society at large, and those of specific subgroups, are ultimately shaped by the competing influences of both types of arguments.

2. Applications. A current example of such a setting pertains to norms of gender relations –say, 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 (personal experiences, high-profile cases, polls, stereotypes) consistent with one view or another circulate, either publicly relayed by the media (probability x) or passed on between people. Some men genuinely care about not harming women (v_H), others are indifferent or misogynistic (v_L), but all want to be seen as being of the first, moral type. The same framework clearly applies to how a dominant national group will "treat," and justify treating, ethnic minorities or immigrants.[14]

Another important case is that of redistribution toward the poor, whether domestic or in the developing world. To what extent are they really suffering and helpless, and conversely how much good (e) does a charitable contribution or a public transfer (if we interpret a as individual tax compliance or voting with the agent taking the composition of public spending as given) really do? Will it make a vital difference to someone's health and their children's education (hence a moral responsibility to act), or is it more likely to be captured by some government or NGO bureaucracy, corrupt local officials, or wasted by the recipients themselves on drugs and alcohol (hence a good reason to abstain)? Another common narrative of the latter kind is that transfers actually harm the poor, by collectively trapping them into a toxic 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 are effective. 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 types but the more moral v_H agents have a stronger influence concern, so they are more inclined to spread positive narratives and refrain from spreading negative ones.

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) For active agents, reputational motives prevail over influence ones whenever the two conflict: when they have a signal indicating that the externality is low, $e = e_-$, both high and low types choose $a_i = 0$ and transmit the excuse to their successor, as in previous sections, even though this may trigger a chain of bad behavior.

[14] In this case and the next, the A and P groups may no longer be of equal size in practice, but this could be adjusted by making the Markov chain (6) non-symmetric. The two groups could also play symmetric roles, with P agents taking another morality-relevant action that affects the welfare of A agents .

(2) In all cases in which they do not know of any narrative (whether learned independently or heard from one's predecessor), high-type agents choose the same *"default action,"* which, with some slight abuse of notation, we shall again denote as $a_H(\emptyset) = 1$ or $a_H(\emptyset) = 0$. As before, we shall analyze both cases in turn.[15]

It will be useful to denote, in either type of equilibrium:

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

$$x_+^P \equiv \Pr\left[i \text{ knows } e \mid i \in P,\ e = e_+\right], \quad x_+^A \equiv \Pr\left[i \text{ knows } e \mid i \in A,\ e = e_+\right]. \tag{9}$$

3.2 Default action of the high type is to act morally

Consider first the case where $a_H(\emptyset) = 1$, meaning that high types always behave prosocially unless they learn of an exculpatory narrative, and conversely observing $a_i = 1$ reveals that they did not receive one. When they do learn of e_- (directly or from their predecessor) all active agents choose $a_i = 0$ and pass on the excuse, since (as stated above) we focus on the case where reputational incentives dominate influence ones. 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 (7), it follows that

$$x_-^P = x + (1-x)(1-\lambda)x_-^A, \quad x_-^A = x + (1-x)\lambda x_-^A, \tag{10}$$

$$x_+^P = x_+^A = x. \tag{11}$$

Consider next the inferences that agents make when their predecessor does not disclose any narrative. There are two main cases to distinguish.

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,

$$\hat{v}_{ND} \equiv E\left[v | a_{i-1} = 0, ND\right] = v_L, \tag{12}$$

$$\hat{e}_{ND} \equiv E\left[e | a_{i-1} = 0, ND\right] = \frac{f_-(1-x_-^A)e_- + f_+ e_+}{f_-(1-x_-^A) + f_+} > e_0. \tag{13}$$

If $i-1$ is an active agent who chose $a_{i-1} = 1$, on the other hand, then he must be a high type, and either $e = e_-$ but he did not know it (otherwise he would have chosen $a_{i-1} = 0$ and disclosed) or else $e = e_+$, in which case he does not disclose, since such signals have neither

[15] Note that depending on whether i's "silent" predecessor was passive or active, and in the latter case on his action choice a_{i-1}, agent i's inferences about e will generally be different, as we shall see. By restricting attention to equilibria in which i takes the same action in these different contingencies, we are thus potentially abstracting from other types, in which these responses differ. This is done to limit the number of cases and focus on the main insights and tradeoffs, which do not depend on this selection.

valuable reputational benefits (given $a_{i-1} = 1$) nor influence on any other high type's actions (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 exactly the same as when observing $(a_{i-1} = 0, ND)$, resulting in the same expected externality $\hat{e}_{ND}$.

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:

$$\tilde{e}_{ND} \equiv E\left[e \mid i-1 \in P, ND\right] = \frac{f_- e_- + f_+(1-x)e_+}{f_- + f_+(1-x)} < e_0. \tag{14}$$

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 trade-offs 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:

$$N_-^A = (1-x)\lambda(\rho + N_-^A) \iff N_-^A = \frac{(1-x)\lambda\rho}{1-(1-x)\lambda}. \tag{15}$$

The conditions for an equilibrium with $a_H(\emptyset) = 1$ can now be written as:

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

$$v_H e_-(1 + N_-^A) - c/\beta \leq \mu(\bar{v} - v_H), \tag{17}$$

$$v_H \tilde{e}_{ND} - c/\beta > \mu\left(v_L - v_H\right), \tag{18}$$

with $\tilde{e}_{ND}$ defined by (14). 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 more others tend to pass them on to others (the higher is N_-^A), the greater will be the social damage from invoking one, 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 $a_H = 1$, which will identify

his type correctly, 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 former case relative to the second: that is why the expected externality involved is $\tilde{e}_{ND} < e_0$ rather than $\hat{e}_{ND} > e_0$.

Proposition 3 (morality as the default behavior). *When (16)-(18) 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 a sharing an excuse is $-e_- N_-^A$, where the virality factor N_-^A is given by (15); such disclosures are therefore strategic substitutes.*

4. *A greater degree of mixing between active and passive agents (lower λ) reduces the multiplier, which simultaneously 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 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.[16]

3.3 Default action of the high type is to act immorally

Consider now the case where $a_H(\emptyset) = 0$, so that high types behave socially only in the presence of a responsibilizing narrative, which they then pass on. This switch to $a = 1$ makes positive-influence concerns relevant for both types –but more so for the high one. In particular, a high type i will now pass on e_+ to $i+1$, as the latter could turn out to be a passive agent (probability $1 - \lambda$) and thus unable to signal through his own actions; being told of the narrative, on the other hand, 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 trade-off: 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

[16] 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,

$$x_-^P = x + (1-x)(1-\lambda)x_-^A, \quad x_-^A = x + (1-x)\lambda x_-^A, \tag{19}$$

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

where the last two equations reflect the fact that if $i-1 \in A$ and knows that $e = e_+$ this information will be transmitted to i only when $i-1$ is a high type. Thus x_-^P and x_-^A are unchanged from the previous case, but x_+^P and x_+^A now have somewhat more complicated expressions (given 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

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

$$N_+^A = (1-x)\left[\lambda\rho(1+N_+^A) + (1-\lambda)N_+^P\right], \tag{22}$$

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

In such an equilibrium, observing an agent who chooses $a_{i-1} = 0$ but provides an excuse e_- leads to attributing to them the pooling reputation $\bar{v}$, since both low and high types behave in this way. In the absence of an excuse, on the other hand, the updated reputation following $a_i = 0$ is

$$\hat{v}_{ND} = \frac{\rho(1-\bar{x}_A)v_H + (1-\rho)[1-f_-x_-^A]v_L}{\rho(1-\bar{x}_A) + (1-\rho)\left[1-f_-x_-^A\right]} < \bar{v}, \tag{23}$$

since if i was of a high type he must have been uninformed (probability $1-\bar{x}_A$), whereas if he was a low type he could either not have had a signal or received and censored e_+ (total probability $1-f_-x_-^A$). As to the expected externality following such an observation, it is

$$\hat{e}_{ND} \equiv E\left[e \mid a_{i-1} = 0, ND\right] = \frac{f_-(1-x_-^A)]e_- + f_+(1-\rho x_+^A)e_+}{f_-(1-x_-^A) + (1-f_-)(1-\rho x_+^A)} > e_0. \tag{24}$$

If the "silent" predecessor $i-1$ was a passive agent, the inference is the same $\tilde{e}_{ND} < e_0$ as before, given by (14). The conditions for an equilibrium with $a_H(\emptyset) = 0$ are then

$$v_H e_- N_-^A = 0 < \mu(\bar{v} - \hat{v}_{ND}), \tag{25}$$

$$v_L e_+ N_+^A < \mu\left(\hat{v}_{ND} - v_L\right), \tag{26}$$

$$c/\beta - v_H\hat{e}_{ND} \geq \mu(v_H - \hat{v}_{ND}), \tag{27}$$

where $\hat{v}_{ND}$ is defined by (23). 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 the responsibilizing narrative 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 would have disclosed it. In contrast to the previous type of equilibrium, the relevant expected externality is here $\hat{e}_{ND} > e_0$ 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.

Proposition 4 (selfishness as the default behavior). *When (26)- (27) 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.4) and (A.5). Such disclosures are therefore strategic complements.*

4. *A greater degree of mixing between active and passive agents (lower λ) lowers N_+^A and raises N_+^P. It simultaneously 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 them if his predecessor $i - 1$ is passive, and similarly after he transmits them to $i + 1$ they are likely to travel further if $i + 1$ is also a passive agent.[17]

3.4 Implications: firewalls, relays and polarization

Note first that the two different types of equilibria and social norms are associated to different circulating narratives. In the "moral" equilibrium ($a_H(\emptyset) = 1$), doing the right thing (e.g., respect toward women) "goes without saying," and conversely deviating requires a justification or rationale, so negative narratives are the ones that will get passed on (when they occur) by

[17] Here again, a lower x increases (both) multipliers, but it now reduces $\bar{e}$; see the Appendix. The intuition is similar to that in footnote 16.

all active agents, and affect behavior. In the "amoral" equilibrium ($a_H(\emptyset) = 0$) self-indulgence is the default, but excuses remain reputationally valuable and thus again circulate. Now, however, so will positive narratives, propagated by passive and high-morality active agents to induce others to behave well.[18]

In either case, Propositions 3-4 show that, within either equilibrium, 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; in the latter case, this also encourages high-morality actors to do the same (strategic complementarity). Through both channels, interspersing them with reputation-concerned actors leads to a "social discourse" and set of beliefs that are on average more moral (or moralizing), as well as *less polarized,* as we show below. Conversely, when active and passive agents –e.g., men and women– interact mostly within segregated pools, very different types of narratives will circulate within each one (mutual stereotyping), with men mostly sharing rationalizations for their behavior, which will be worse on average than under integration, and women mostly judging it as inexcusable.

Proposition 5 (polarization of narratives and beliefs)**.** *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* $\lambda = 1$.[19]

4 Moral Standards: What is Justifiable?

4.1 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

[18] Here again, both equilibria may coexist for some range of parameters, but Pareto-dominance now has 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, and may not even be alive at the same time.

[19] For the equilibrium of Proposition 3, one of the U-shapes is degenerate, in that $\left|x_+^P/x_+^A - 1\right|$ is equal to zero for all λ. For $\left|x_-^P/x_-^A - 1\right|$ the monotonicities on either side of $1/2$ are strict and symmetric, so the global maximum is also reached at $\lambda = 0$. For the equilibrium of Proposition 4, all statements hold in the strict sense.

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.[20] 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

$$\mathcal{M}^-(e) \equiv E_F\left[\tilde{e} \mid \tilde{e} < e\right] \quad \text{and} \quad \mathcal{M}^+(e) \equiv E_F\left[\tilde{e} \mid \tilde{e} \geq e\right], \tag{28}$$

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

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

Moral standards. We shall focus attention on equilibria taking the following intuitive form: when the signal e about the importance of the externality is below some cutoff $\hat{e}$, both types disclose this "excuse" and choose $a = 0$; when it is above, the high type chooses $a = 1$, perfectly separating himself, and neither type discloses e (as this would be useless for the high type, and self-incriminating for the low one).

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

$$\hat{v}_D = \frac{\rho x_H v_H + (1-\rho) x_L v_L}{\rho x_H + (1-\rho) x_L}, \tag{29}$$

[20] 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).

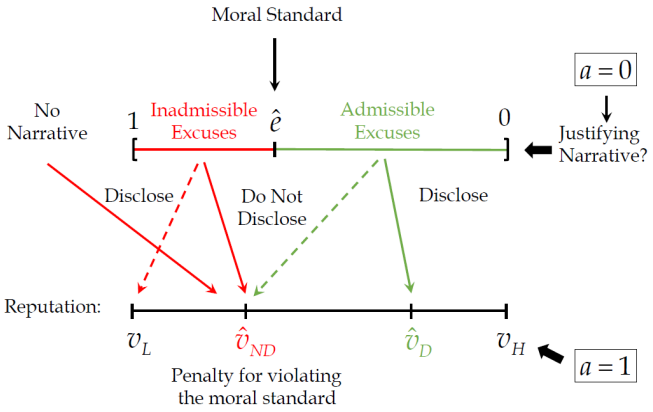

Figure 2: Moral Standards and Narratives. Straight arrows describe equilibrium play, dashed ones off-path deviations

which is independent of e.[21] 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:

$$v_H \hat{e} - c/\beta + \mu(v_H - \hat{v}_D) \equiv 0. \tag{30}$$

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 $\hat{v}_{ND} < \hat{v}_D$.[22]

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 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 (30) 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.

[21] The denominator is always well-defined, as there is no equilibrium (in undominated strategies) in which $(x_H, x_L) = (0, 0)$; see the Supplementary 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.

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

4.2 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 $v_H e_0 - c/\beta + \mu(v_H - v_L) \geq 0$, or substituting in (3):

$$v_H(e_0 - e^*) \geq \mu(v_L - \bar{v}) = -\mu\rho(v_H - v_L). \tag{31}$$

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 (30), 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 $v_H e - c/\beta + \mu v_H$.[23] His expected utility from a search intensity x is therefore

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

leading to the first-order condition

$$\psi'(x_H) = F(\hat{e})\left[c/\beta - \mu(v_H - \hat{v}_D) - v_H\mathcal{M}^-(\hat{e})\right] = F(\hat{e})v_H\left[\hat{e} - \mathcal{M}^-(\hat{e})\right]. \tag{32}$$

The low type, trying to mimic the high one, will only disclose those same values $e < \hat{e}$, 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

$$\psi'(x_L) = \mu F(\hat{e})(\hat{v}_D - v_L). \tag{33}$$

[23] 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 (32), (35) and (38) 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 6 this makes equilibria more likely to be of the type where $x_H > x_L$ than the reverse, and in Proposition 8, where only the first case is possible, it helps sustain the existence of such an equilibrium.

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 (29)-(33), together with a prior e_0 high enough that (31) holds. Furthermore, $x_H > x_L$ if and only if $\mathcal{M}^-(\hat{e})v_H \leq c/\beta - \mu(v_H - v_L)$ or, equivalently

$$\mathcal{M}^-(\hat{e})v_H \leq c/\beta - \mu(v_H - v_L). \tag{34}$$

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 (32) 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 6 (prosocial norm). *For any e_0 high enough that (31) 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 $e < \hat{e}$) 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

$$e^* - \mathcal{M}^-(e^*) > \mu\rho(v_H - v_L)/v_H, \tag{35}$$

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 (35) is reversed, then 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 (35), 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 7. *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 (35) 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 (30)), 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.

4.3 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 (29) and the same threshold $\hat{e}$ as in (30). Beliefs following $(a = 0, ND)$, however, are now

$$\hat{v}_{ND} = \frac{\rho\,(1 - x_H)\,v_H + (1-\rho)[1 - x_L F(\hat{e})]v_L}{\rho\,(1 - x_H) + (1-\rho)[1 - x_L F(\hat{e})]} > v_L. \tag{36}$$

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 $v_H e_0 - c/\beta + \mu(v_H - \hat{v}_{ND}) \leq 0$ or, equivalently

$$v_H(e_0 - e^*) \leq \mu(\hat{v}_{ND} - \bar{v}). \tag{37}$$

As expected, this now puts an upper bound on the prior e_0 about the severity of the externality

($e_0 v_H \leq c/\beta$). Conversely, even though $\hat{v}_{ND}$ depends on the distribution F and thus on its mean e_0, (37) 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):

$$\psi'(x_H) = \mu\left(\hat{v}_D - \hat{v}_{ND}\right) + \left[1 - F(\hat{e})\right]\left[\mathcal{M}^+(\hat{e}) - \hat{e}\right]v_H, \tag{38}$$

$$\psi'(x_L) = \mu F(\hat{e})\left(\hat{v}_D - \hat{v}_{ND}\right). \tag{39}$$

so it must always be that $x_H > x_L$, which as noted earlier implies that $\hat{v}_D > \bar{v}$ and $\hat{e} > e^*$.[24]

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 (29) and (38)-(39), together with a prior e_0 low enough for (37) 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 8 (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, (37) holds) and violating the moral standard (behaving selfishly without a narrative $e < \hat{e}$) 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 (31) and (37) can be shown to be compatible over a range of priors, so that both types of equilibria can coexist.

Proposition 9 (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 ($\bar{v} > \hat{v}_D > \hat{v}_{ND} = v_L$).

(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.

[24] Clearly, $x_H \geq x_L$. Equality would mean that $\hat{v}_D = \bar{v}$ and hence $\hat{e} = e^* < 1$, which given full support of f would imply that $F(\hat{e}) < 1$ and $\mathcal{M}^+(\hat{e}) > \hat{e}$; (38) would then lead to $x_H > x_L$, a contradiction.

(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).

5 Narratives Versus Imperatives

Having enriched the reputation or identity channel in the previous section, we now expand the influence channel, focusing again on interactions within a single dyad. One actor is now a principal (she) whose only decision is how to communicate with an agent (he), who will in turn takes 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* refers to a recommended action –say, to do $a = 1$. It does not focus on motives, but on the decision itself; it is a precept to behave in a certain way. Both narratives and imperatives are commonly used to instill moral behavior. Many scholars have argued that oral, written, and more recently cinematic narratives are essential components of "effective moral education" (Vitz 1990, p. 709).[25] Imperatives, in contrast, do not take the form of stories: they are commands of the form "Thou shalt not kill," as codified for instance in the Ten Commandments. This notion is reminiscent of deontological or rule-based moral reasoning, put forward by Immanuel Kant. An essential difference between utilitarian/consequentialist versus deontological normative theories is that, in the former, only consequences (life, happiness, welfare) conceivably justify moral decision making, whereas the latter postulates a categorical demand or prohibition of actions, no matter how morally good their consequences: what makes a choice right is its conformity

[25] See also the experimental-economics literature on moral persuasion, e.g., Dal Bó and Dal Bó (2014).

with a moral imperative.[26]

2. Evidence. Empirically, behavior seems to reflect a mix of utilitarian and deontological behaviors –both between and within individuals. In the trolley or transplant problems mentioned earlier, consequentialist reasoning requires "actively" killing one to save many, while deontological reasoning calls for obeying the imperative not to kill, regardless of consequences and without reference to any ends. Faced with such dilemmas, different people choose differently, and even minor variations in framing can drastically alter their (hypothetical) decisions. Likewise, no consensus is reached concerning so-called "repugnant goods" (Roth 2007). While advocates of markets for organs, sex or surrogate motherhood point to the large potential welfare gains, among the opponents are some who give priority over these benefits to the imperative never to treat human beings as a means to an end.

Experiments with real stakes, in which subjects choose between money and a charitable act (c versus e in our notation) under varying probabilities that their decision will actually be implemented, paint a similar picture. Feddersen and Sandroni (2009) and Chen and Schonger (2013) note that if the same probability applies to costs and benefits, the behaviors of both pure consequentialists (in the traditional sense that excludes any image concerns) and of people with purely expressive or/and deontological preference should both be entirely invariant to it. In practice it does matter (positively), which reveals a tradeoff between the two types of motives. In Falk and Szech (2017), a group of subjects vote simultaneously on killing a mouse, or on destroying a charitable donation, in return for money. In both paradigms (mice and donation), the decision over c is now individual and non-contingent, whereas the probability that it can bring about e, which requires one's vote to be pivotal, varies. The frequency of moral choices is found to decrease as the chance of making a difference falls, in line with consequentialism, but even as it reaches zero there remain about 18% of subjects (in both cases) who will not take the money "on principle." In actual elections, similarly, substantial fractions of people participate in spite of not being pivotal, presumably an expression of a moral or civic "duty".

3. 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 (as those in Section 4, or below as potential misinterpretations of finer-grained messages), imperatives provide quick, instinctive guides to decisions in unfamiliar contexts where moral consequences may be complex to figure out (e.g., trolley problems). Because they embody only coarse information, on the other hand, they will sometimes induce moral mistakes, at least from a utilitarian point of view. This may occur even in the presence of relatively precise signals about consequences (e in the model), if the source of that narrative is deemed less reliable than the source of a preexisting imperative, and thus has a weaker effect on beliefs.

Most puzzling are situations in which the agent is fully aware of consequences –has the right posterior beliefs– and makes a reflective (not instinctive) decision to ignore them in favor of an

[26] The discussion about these two main lines of thought in moral philosophy is, of course, much 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.

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 actually benefiting any recipient. This inherent rigidity of imperatives is connected to the rules vs. discretion idea and 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 (from an ex-ante point of view) 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," as in Bénabou and Tirole (2004). When made-up excuses and rationalizing narratives 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.[27]

The formal analysis will confirm the importance of the key factors, foreshadowed above, that are conducive to imperatives being issued and obeyed.[28] These factors stem from both demand and supply, i.e., characteristics of agent and principal.[29] First, imperatives are effective only if issued by highly trusted principals, whereas everyone can use the narrative route to attempt to persuade. Second, imperatives coarse and cheap-talk nature economizes on cognitive costs and makes them less fragile to interpretation uncertainty (whether accidental or self-serving) 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 particularly valuable to deal with *temptations.* Fourth, and relatedly, by pooling states in which the agent would be reluctant to behave in the recommended way with others where he would be willing, imperatives expand the scope for the desired behavior.

5.1 Modeling imperatives

There is a principal (she) who learns a narrative drawn according to a continuous $F(e)$, and an agent (he) who does not and will choose an action $a = 0, 1$. There may also be a passive

[27] We model in the Appendix a related phenomenon, namely the fact that 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.

[28] Roemer (2010) offers a different approach to modeling agents obeying the categorical imperative: a strategy profile constitutes a Kantian equilibrium if no individual would like all players to alter their contributions by the same multiplicative factor. Studying evolutionary stability of moral preferences, Alger and Weibull (2013) show that a combination of selfish and "Kantian" preferences (now in the sense of choosing to do "the right thing" provided all others do as well) is stable in equilibrium.

[29] 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).

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.[30]

1. Preferences. Let $U^A(e)$ and $U^P(e)$ respectively denote the moral agent's and the principal's net returns from the former's 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, whereas the principal internalizes their externalities on a larger group (e.g., whole population vs. in-group only), or their "internalities" due to imperfect self-control, or derives other private benefits from the choice of $a = 1$. We then have

$$U^A(e) = v_H e - c/\beta + \mu(v_H - \bar{v}) = v_H(e - e^*) \tag{40}$$

for the (high-type) agent at the moment of decision, while for the principal, who takes an ex-ante and more inclusive perspective,

$$U^P(e) = E_{\tilde{v}}\left[[we + (\tilde{v}e - c)]\, a(\tilde{v})\right] \equiv \rho\left(we + v_H e - c\right), \tag{41}$$

where $w \geq 0$ measures the extra value she attaches to moral behavior by the agents, on top of his expected welfare. Her indifference point, $e^P \equiv c/(w + v_H)$, will be lower relative to the high-type agent's e^*, the higher is w, the more present-biased and the less image-conscious he is (as reputation-seeking is a zero-sum game). Thus, the greater the gap $e^* - e^P$, for instance due to low self-control β, the greater the potential role for imperatives and positive narratives. The case $w = 0$ corresponds to a sophisticated individual's ex-ante self (or parents maximizing their child's ex-ante welfare), that of $w = 1$ to a utilitarian social planner, and that of $w = +\infty$ to a passive actor wanting to promote the action $a = 1$ but having no empathy for the individual.

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 (the closer is e^P to e^*, the better aligned are their interests) and assume that the agent knows $U^P(\cdot)$ –that is, how trustworthy the principal is.

2. Communication. When she has a convincing narrative $e > e^*$, the principal can disclose it to the agent. Whether or not she has a narrative of any kind, she can instead, or also, issue an imperative. The first form of communication consists in laying out, whether through facts or by telling a story, specific reasons why the agent should choose that course of action. The second one amounts to telling him "do it just because I said so," or "just trust me."[31] Kant's

[30] If the Principal always wanted $a = 1$ no credible imperative could ever be issued, which is why imperatives were not considered in Section 3.

[31] As we shall see, in equilibrium the principal will not use both routes simultaneously. See Dewatripont and Tirole (2005) for an analysis of rival modes of communication in a different context.

categorical imperative to "Act only on that maxim through which you can at the same time will that it should become a universal law" can also be understood in this light: by visually scaling up, in a recipients' mind, the externalities attached to his actions, it operates as a prescription to always act as if the consequences of one's action were very large, without probing into their actual magnitude.

5.2 Coarse versus noisy communication

We develop here the idea that, while imperatives are a more coarse and "cheap-talk" form of information than narratives (which may even involve actual evidence), their advantage is that they are more easily communicated and understood. Such cheap-talk messages only work, however, if they are credible enough to raise the agent's posterior above e^*, thus persuading him to indeed choose $a = 1$.

In the absence of imperatives, the natural equilibrium would be for the principal to communicate all narratives $e \geq e^*$ and say nothing otherwise, a silence which a rational agent would then correctly interpret as meaning that $e < e^*$. Suppose now that when the principal tries to convey an argument $e \geq e^*$, there is some small probability $1-\xi$ that the agent does not receive the message (e.g., did not hear it, was not paying attention), or cannot make sense of it except as random noise. In such cases he will revise his belief about e downward from $e_0 < e^*$ rather than upward, especially if $1 - \xi$ is small, and 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

$$\mathcal{M}^+(e^P) \equiv E[e \mid e \geq e^P] \geq e^*. \tag{42}$$

When (42) 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 (42) fails; the principal uses narratives instead, when she has them, and compliance is only $\xi[1 - F(e^*)] < 1 - F(e^P)$. Condition (42) 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 $F_1(e)$ is second-order stochastically dominated by $F_2(e)$. By Lemma 1, 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 having better judgment confers greater "moral standing" to a principal, allowing her to more credibly issue imperatives: (42) becomes more likely to hold. Narratives, in contrast, can be spread by everyone.[32]

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 $\theta_1 < \theta_2$:[33]

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

Proposition 10 (**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.[34] 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 education processes combine a variety of narratives and imperatives; for instance, a story about someone who has

[32] 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 2.4, 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.).

[33] 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 $\mathcal{M}^+$.

[34] 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 (for models of the latter see, e.g., Bisin and Verdier (2001) or Tabellini (2008).

stolen or lied and came to regret it, followed by a generalization to "thou shalt not steal/lie." Providing a general treatment of the narrative-imperative mix lies outside the scope of this paper –requiring assumptions on how the two types of messages combine to generate updated beliefs, the relative likelihood that each of them is understood or trusted by the agent, etc. Here we simply outline a stylized example showing how, when congruence is too small to generate a credible imperative on a stand-alone basis, the principal may start with some narrative(s) that raise her credibility enough that an imperative then becomes effective.

Let $\hat{e} < e^*$ be defined by $\mathcal{M}^+(\hat{e}) = e^*$, and suppose that $e^P < \hat{e}$, so that (42) 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 will in turn render credible issuing the imperative "do $a = 1$" for all values $e > e^P$, whereas on its own it would not be.

5.3 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 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.

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 (ii), 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 all narratives e,[35] 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

[35] Even with a small disclosure cost the principal will now reveal realizations $e < e^*$, since $\Pr[\varepsilon > e^*|e] > 0$.

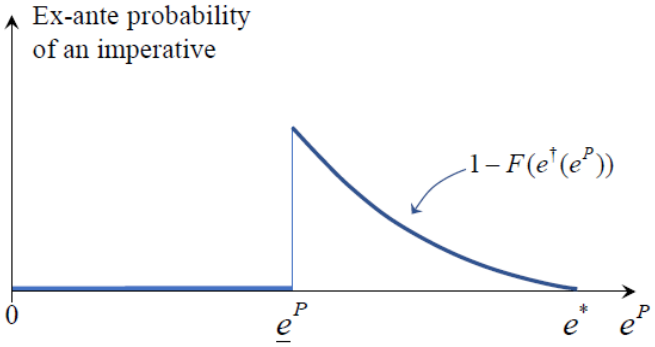

Figure 3: Use of an Imperative

(effective) imperative he would have chosen $a = 1$. Equilibrium behavior therefore requires that

$$\Delta(e) \equiv \int_0^{e^*} U^P(\varepsilon) dH(\varepsilon|e) \geq 0 \quad \text{for all } e \in I, \tag{43}$$

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 (43) is never satisfied at $e = e^P$, since

$$\Delta(e^P) + \int_{e^*}^1 U^P(\varepsilon) dH(\varepsilon|e^P) = \int_0^1 U^P(\varepsilon) dH(\varepsilon|e^P) = U^P(e^P) \equiv 0, \tag{44}$$

where the second 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 $e_2 > e_1$ (this is shown in the Appendix). Therefore, I must be some interval 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

$$\mathcal{M}^+(e^{\dagger}) \geq e^*, \quad \text{where} \quad \Delta(e^{\dagger}) \equiv 0. \tag{45}$$

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 (45) 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 $1 - F(e^{\dagger})$ of the latter decreases toward 0; see Figure 3.[36]

2. Self-control. The analysis, applied with the benchmark preferences (40)-(41), reveals a similar ambiguous impact of self-control. On the one hand, as β decreases the principal is more tempted to use an imperative (e^* increases, and so $e^{\dagger}$, given by setting (43) to zero, decreases: I expands). On the other hand, at some point the obedience condition $\mathcal{M}^+[e^{\dagger}(e^*)] \geq e^*$ may no longer be satisfied, as $\mathcal{M}^+(e^{\dagger})$ decreases and e^* increases. A worsening self-control problem facilitates the emergence of an imperative (or rigid personal rule), but only up to a point, as

[36] 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 $\mathcal{M}^+(e^{\dagger}(\underline{e}^P)) = e^*$.

shown in Figure 3 (replacing e^P by β).

Proposition 11 (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 (43).*

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 issued is hump-shaped in congruence: zero below some minimum level, then positive but decreasing back toward zero as congruence increases further toward perfect alignment of interests. 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.

6 Concluding Remarks

We have developed here a simple but fairly comprehensive framework to analyze moral behavior and its malleability. Together with known key 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 organize and interpret many empirical findings, and generates new predictions.

Of course, the analysis still leaves many important issues to be explored in future research. First, we have modelled 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 have taken as a primitive some class of arguments that "work" in persuading agents about the magnitude of externalities, and focused on analyzing how people will then *search* for them, *invoke* them, *repeat* them, and *judge* those who do so. What makes so many "content-free" narratives work is still imperfectly understood, however, beyond the fact that it typically involves important elements of both heuristic and motivated or wishful thinking –models of which could be combined with the present one.

An important related question is that of competing narratives. We focussed on a single moral argument that each person may or may not be exposed to and use, but in reality parties

with conflicting interests, or opportunistic "narrative entrepreneurs," will offer very different rationales for what is right or wrong. What factors then make one story more compelling than the other? Related areas of potential investigation include politics (fake news, focus of the debate), the construction of identity and identity conflicts (in-group/out-group narratives), and the design of religious and other doctrines (complementarity or substitutability between narratives and imperatives).

Disagreements over what constitutes a moral act can also extend beyond differences in 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 (externality) conception and suggests the inclusion of other phenomena such as loyalty, authority, and spiritual purity. These notions can, in large part, already be mapped back to our model through the (self) signaling of personal values and of the in-group they extend to, but working out more specific and testable implications (e.g., where a tradeoff arises not with some private cost but with a concern about harming others) seems worthwhile.

Differing social preferences even under full-information (or, alternatively, heterogenous priors) constitute another potential source of disagreement, often relevant for "hot" societal issues such as abortion, gay rights, gun control, etc. If agreement on what constitutes a negative versus positive externality is not commonly shared, social image becomes multidimensional and group-dependent. Our model abstracts from these tradeoffs by implicitly assuming that people care only (more generally: on net) for the esteem of those who share their basic values, but reputational concerns are often more complex and wide-ranging.

Finally, a more methodological direction that we start exploring in a companion paper (Bénabou, Falk and Tirole 2018) is the extent to which a person's true moral preferences can be retrieved from observing their choices over a broad range of situations, and how different experimental methodologies perform in that respect. This also has implications for how one should interpret Kantian-like choices (refusals of tradeoffs) that appear to be deontologically rather than consequentially motivated, and for potential improvements in contingent-valuations methods and other measures of willingness to pay for public goods.

Appendix: Main Proofs

Proof of Proposition 3 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 (10)) 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 4 We first solve the system (20) to obtain

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

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

where

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

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

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

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

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

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

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 (20)) of being informed of e_+ when it occurs, in which case the high type will switch to $a=1$; therefore, $\bar{e} = (\rho/2)f_+ e_+ x_+^A$. The formula for x_+^A derived above shows that it

is a rational fraction in λ, with determinant equal to $(1-x)$ times

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

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

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

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 5 For negative signals, x_-^P and x_-^A are given by (19), 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

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

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

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

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

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

completes the proof. ■

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

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

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

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

(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 6 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 (30), while (32)-(33) imply that $x_L \leq x_H$ if an only if:

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

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

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 $\mathcal{M}^-(e^*)v_H > c/\beta - \mu(v_H - v_L)$, which involves only exogenous parameters. It will hold if F is insufficiently bottom-heavy, or too top-heavy.[37] Rewriting the inequality slightly using (3) yields the reverse of (35).

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

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

which will hold when F is sufficiently top-heavy ($\mathcal{M}^-(\hat{e})$ close to $\hat{e}$, meaning that F has relatively little mass below $\hat{e}$), or more generally not too bottom-heavy (which would make $\mathcal{M}^-(\hat{e})$ close to zero). In particular, a *sufficient* condition on exogenous parameters that *precludes* any such equilibrium is $\mathcal{M}^-(e^*)v_H < c/\beta - \mu(v_H - v_L)$, which holds when F is insufficiently top-heavy, or too bottom-heavy. Rewriting the inequality slightly using (3) yields (35).

It only remains to prove that an equilibrium with $a_H(\emptyset) = 1$ exists whenever (31) is satisfied. Equation (30) 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 (32)-(33) then associate a unique pair $(x_H, x_L) \in [0, 1]^2$, 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, (29) associates a new value $\hat{v}'_D \in [v_L, v_H]$. Moreover, 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 8 Each type's expected utility from a search intensity x are now

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

leading to the stated first-order conditions. It remains to prove that an equilibrium with $a_H(\emptyset) = 0$ exists when e_0 is low enough. Equation (30) 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 (38)-(39) 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, (36) 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 $\hat{v}_{ND} = \hat{v}'_{ND}$

[37] 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).

in $[v_L, \bar{v})$. For $v_H(e_0 - e^*) < \mu(v_L - \bar{v}) = -\mu\rho(v_H - v_L)$, moreover, equation (37) must then hold, so all equilibrium conditions are satisfied. ■

Proof of Proposition 9 Let $\underline{e}_0$ be the value of e_0 that makes (31) 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 (30), (38)-(39) and (36); we saw in the proof of Proposition 8 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 (37). 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 (36), it suffices that $x_H(e_0)$ remain bounded away from 1, and from (38) this is ensured as long as $\psi'(1) = +\infty$, since the right-hand side of (38) is bounded above by $\mu(v_H - v_L) + v_H[\mathcal{M}^+(\hat{e}) - \hat{e}] < \mu(v_H - v_L) + v_H$. ■

Proof of Proposition 10 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 11 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 10. (Recall that the payoff when disclosing a narrative is equilibrium-independent).

The only claims remaining to prove are equation (44) 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,

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

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

Supplement to Section 5: 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:

$$v_2 e^P - \frac{c}{\beta} \geq 0.$$

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 $\hat{v} = v_1$.[38]

For type v_1 to obey the imperative in such an equilibriums, it must be that:[39]

$$\int_{e^P}^{1} \left(\frac{c}{\beta} - v_1 e \right) \frac{dF(e)}{1 - F(e^P)} \leq \mu \left[v - \frac{\rho(1-\lambda) v_1}{1 - \rho\lambda} \right]. \tag{A.9}$$

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

$$\int_{e^P}^{\frac{c}{\beta v_1}} \left(\frac{c}{\beta} - v_1 e \right) \frac{dF(e)}{1 - F(e^P)} < \mu (v - v_1) = \mu\lambda (v_2 - v_1). \tag{A.10}$$

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.9) exceeds that of (A.10). Because the left-hand side (A.10) exceeds that of (A.9), condition (A.9) is verified if (A.10) is. Hence:

Provided that (A.10) 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.

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

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

Supplementary Appendix: 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:*

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

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

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

Furthermore, the following three properties are equivalent:

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

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

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

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

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

and observe from (B.2) that $\hat{L}_{ND} = Q(P(N))$. It is easily verified that Q is increasing in Z if condition (ii) holds, and decreasing when it is reversed. Note also, from (B.1), 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-equilbrium beliefs, moreover, the high type may prefer to choose $a = 1$ even at relatively low values of e, meaning that his equilibrium choice of a is no longer a threshold rule.

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

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

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

Proposition 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:*

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

Under either condition in Claim 3 $\hat{v}_D > \hat{v}_{ND}$, so this reduces to (30), 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. ■

[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.

References

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.

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

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.

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

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.

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.

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., Fehr, E., and D. Schunk (2012) "Health Effects on Children's Willingness to Compete," *Experimental Economics*, 15(1): 58–70.

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.

Baumeister, R. F., Vohs, K. D., and D. M. Tice (2007) "The Strength Model of Self-Control," *Current Directions in Psychological Science*, 16(6): 351–355.

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.(2006)

_____ (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, 0(3), 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.

Chen, D. L. and M. Schonger (2013) "Social Preferences or Sacred Values? Theory and Evidence of Deontological Motivations," Discussion Paper, ETH Zurich.

Clot, S., G. Grolleau, and L. Ibanez (2018). "Shall We Pay All? An Experimental Test of Random Incentivized Systems," *Journal of Behavioral and Experimental Economics*, forthcoming.

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.

Dewatripont, M. and J. Tirole (2005) "Modes of Communication," *Journal of Political Economy*, 113: 1217–1238. Diamond, P. and J. Hausman (1994) "Contingent Valuation: Is Some Number Better than No Number?," *Journal of Economic Perspectives,* 8(4): 45–64.

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, pp. 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.

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.

Feddersen, T., Gailmard, S. and A. Sandroni (2009) "Moral Bias in Large Elections: Theory and Experimental Evidence," *American Political Science Review*, 103(2): 175-192.

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

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.

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

Golub, B. and E. Sadler (2016) "Learning in Social Networks," In Y. Bramoullé, A. Galeotti, and B. W. Rogers (eds.), The Oxford Handbook of the Economics of Networks, Chapter 19: 504–542. New York: Oxford University Press.

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.

Haley, K. J. and D. M. T. Fessler (2005) "Nobody's Watching? Subtle Cues Affect Generosity in an Anonymous Economic Game," *Evolution and Human Behavior* 26: 245-256.

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) *The Handbook of Experimental Economics*. Princeton, Princeton University Press.

Kahan, D. (2013) "Ideology, Motivated Reasoning, and Cognitive Reflection," *Judgment and Decision Making*, 8: 407–424.

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.

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).

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.

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,

Roth, A. E. (2007) "Repugnance as a Constraint on Markets," *Journal of Economic Perspectives*, 21(3): 37–58.

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

Sandel, M. (2010) *Justice: What's the Right Thing to Do?* New York: Farrar, Strauss and Giroux.

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

Tversky, A. and D. Kahneman (1981) "The Framing of Decisions and the Psychology of Choice," *Science*, 211(4481): 453–458.

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.