Narratives, Imperatives and Moral Reasoning

Roland Bénabou, Armin Falk, and Jean Tirole

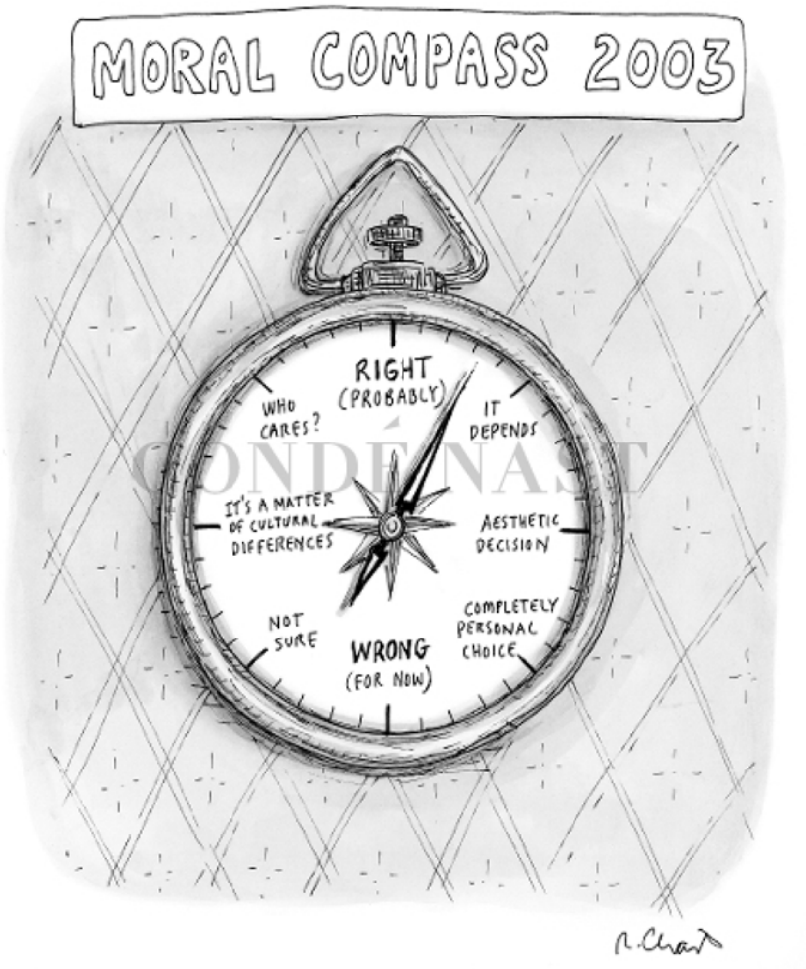
Princeton University – Bonn University – Toulouse School of Economics

April 2019

What is the moral thing to do?

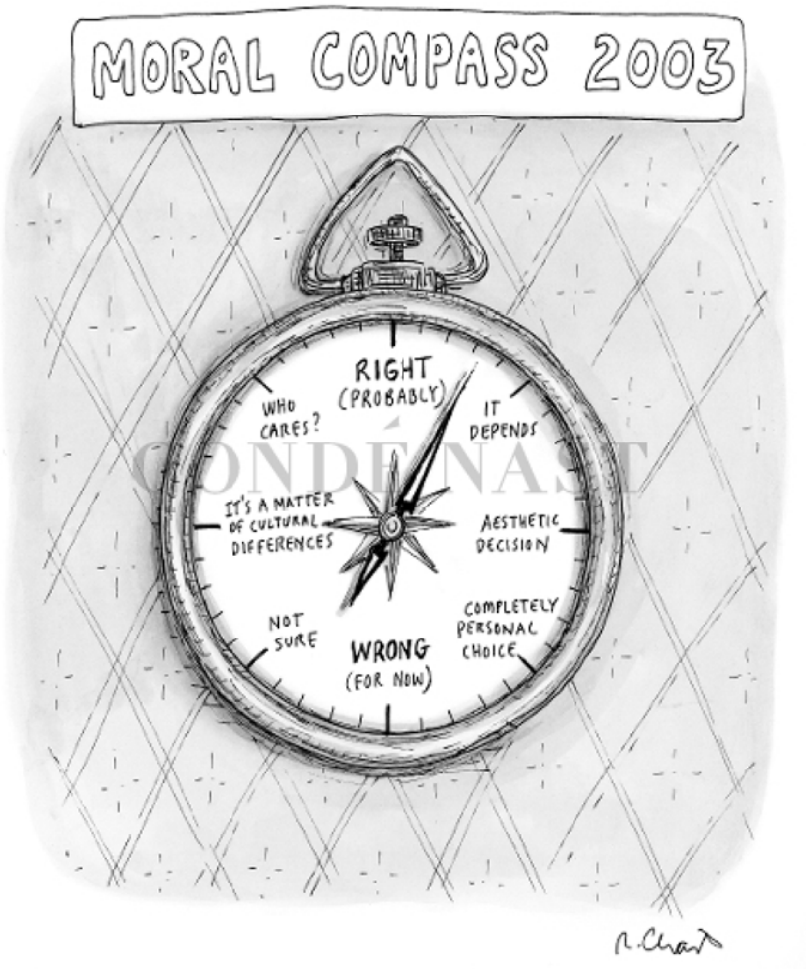
- Will of not aim to answer that question, but instead to analyze the production and circulation of arguments seeking to justify one or another course of action on the basis of morality
  - Such appeals to notions of "right or wrong" pervade social and political discourse, often trumping any argument of economic efficiency (banning "immoral" transactions, trade wars, (un)deservedness of some group...)
- Two main types of moral arguments:
  - Provide reasons for what one "ought to do," or on the contrary justifications for acting according to self-interest, under specific circumstances: narratives
  - Broad "fiat" prescriptions, dictating a fixed behavior across most settings, without explaining why: imperatives

Narratives...



Narratives

... and Imperatives



Narratives

Imperatives

Outline

1. Start from workhorse model capturing basic determinants & regularities of moral behavior: intrinsic values + incentives + self/social image or identity [B-T 2011a,b]. Next: combine costly actions + communication about reasons
   - Introduce moral narratives or rationales, their sources and functions:

     "Excuses" $\leftrightarrow$ reputation concerns, "responsibilities" $\leftrightarrow$ influence concerns
2. Social transmission / virality of narratives
   - Embed this into model of strategic communication on heterogenous network
   - When will exculpatory versus responsibilizing narratives spread or remain clustered? Resulting norms of morality?
3. Moral standards
   - Allow endogenous search for reasons to act / not act prosocially
   - How strong do excuses for selfish behavior need to be in order to be acceptable? How does society judge those who do or do not have one?
4. Narratives and imperatives as means of moral influence
   - Imperatives' clarity vs. credibility; what confers someone "moral legitimacy" to issue them and be obeyed? Flexibility costs.

Related literature

1. Morality, prices and markets.
   - [Brekke et al. 2003, Roemer, J. 2010, Falk and Szech 2013, 2014, Elias et al. 2016, Ambuehl 2015...]
   - Prosocial behavior and signaling, moral identity [Bénabou and Tirole 2006, 2011, Dana et al. 2007, Ellingsen and Johannesson 2008, Ariely et al. 2009, Exley 2016, DellaVigna et al. 2016, Gino et al. 2016, Grossman and van der Weele 2017...]
2. Public goods and learning in networks
   - Probabilistic contagion and/or non-strategic communication
   - Strategic communication [Hagenbach and Koessler 2010, Galeotti et al. 2013, Ambrus et al. 2013, Acemoglu and Jackson 2015, Bloch et al. 2016, Foerster and van der Weele 2018.]
3. Culture and its transmission
   - Values, Beliefs [Bisin, A., and T. Verdier 2001, Tabellini 2008, Bénabou and Tirole 2006b, Dohmen et al. 2012, Besley and Persson 2016]
   - Narratives, memes, folklore [Shiller 2017, Mukand and Rodrik 2016, Barrera et al. 2017, Michalopoulos and Meng Xue 2018...]

1. Basic Model of Moral Behavior

- Individuals may engage in moral/pro-social behavior $(a = 1)$ or not $(a = 0)$
- Cost c, perceived as c/β where $\beta \leq 1$ is self-control parameter.
- Social externality: perceived magnitude (or probability) $e \in [0, 1]$
- Consequentialism: agents have intrinsic motivation / moral values $v \cdot e$
- Types:
  - High or "moral" type: $v = v_H$, probability ρ
  - Low or "immoral" type: $v = v_L$, probability $1 - \rho$
  - Mean: $\bar{v} \equiv \rho v_H + (1 - \rho) v_L$
- Will generally ensure low type never contributes, focus on high type's behavior

Self or social image motive

- In addition to intrinsic motivation ve, cost or extrinsic incentives c, agents care about social and/or self-esteem

$$U = \left(ve - \frac{c}{\beta} \right) a + \mu \hat{v}(a) \qquad \text{for } v = v_H \text{ or } v = v_L$$

- $\hat{v} = E[v \mid a; c, \mu, ...]$: perceptions of individual's "true values"
  - Intrinsic or instrumental motive: predictor of what is likely to do in the future, especially absent incentives, when no one is looking, etc.
- μ = strength of (self or social) image concerns. Same for all here
  - Also, agreement in reference group /audience on what makes a positive / negative externality $\Rightarrow$ on higher v's being more socially desirable. Can relax

Assumption

$$v_H - \frac{c}{\beta} + \mu \underbrace{(v_H - \bar{v})}_{\text{rep. gain when } a_H = a_L = 0} > 0 > v_L - \frac{c}{\beta} + \mu \underbrace{(v_H - v_L)}_{\text{max. rep. gain}}$$

- Implies low type never contributes (dominant strategy)

Basic determinants of moral behavior

Proposition

The moral type contributes iff $e > e^*$, *where* e^* *is uniquely defined by*

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

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

- Substantial experimental / empirical evidence supporting these predictions
- Selfish behavior facilitated by good initial reputation ρ: "moral licensing"
  - More generally, hump-shape (BT 2011): both moral licensing & "foot-in-the-door"
- Add: key role of beliefs (actors', observers') and strategic communication about moral consequences and meaning of actions (e.g., level of e)
  - Paper considers: disclosure, search, cheap talk

2. Introducing Narratives

- Any news, story, experience, rationalization, heuristic, that can potentially alter agents' beliefs about tradeoff between private benefits and social costs (or vice versa) faced by a decision-maker. This actor could be agent himself, someone he observes, or someone he seeks to influence
- Formalize as persuasive signal about importance of externality e
  - Can also be about cost c, visibility/salience μ
  - Acts as hard information (with prob > 0). Treat as such, though need not be
- What matters is having perceived "grain of truth, sense-making":
  - True, relevant fact or causal relationship: pollution, poverty...
  - Negative group stereotypes, boys will be boys, "alternative facts," ...
  - Pure frame, euphemisms: pro-life, pro-choice, collateral damage...
  - All of the above: court argument
- Many bounded-rationality channels may be involved
  - Memory: retrieval is selective, cue-activated. Overweighing: base-rate neglect, salience, simplicity/vividness, emotional charge. Motivated cognition. Correlation/causality...
- Veracity / accuracy does not matter for positive analysis (our main focus)

Types of moral narratives

1. Negative narrative (excuse, absolving): pushes toward $a = 0$
   - Downplay externality: minimize harm, blame the victim, dehumanizing language: undeserving poor, Nazi propaganda on Jews, etc.
   - Low level of pivotality: bystander effect, large groups, markets: "If I don't do it, someone else will". Consequentialist/utilitarian reasoning.
   - Magnification of personal cost: "We only followed orders," "I was going through a hard time, wasn't thinking," "I am only doing my job",...
2. Positive narrative (duty, responsibility): pushes toward $a = 1$
   - Moral precepts, religion, fairy tales, heroes, role models...
   - Imaginary counterfactuals: "What if everyone did that? Related to Kant's categorical imperative
   - Images and cues inducing empathy ("What if it were you?") and/or stressing common identities (nation, brotherhood, loyalty, same planet)

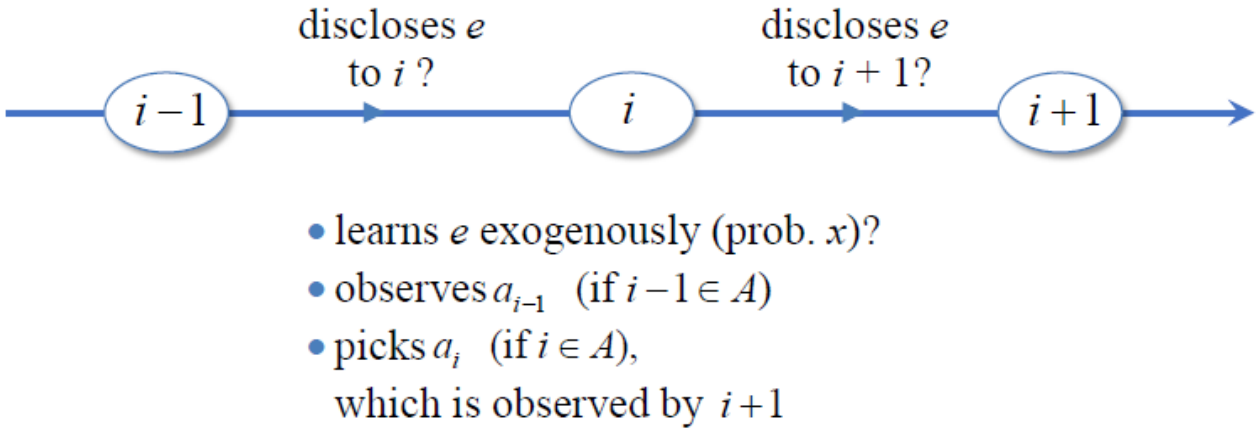
- Formalize as: externality initially unknown, and represent
  - any narrative that might be obtained $\leftrightarrow$ signal inducing some posterior belief distributed a priori as $F(e)$ on $[0, 1]$

3. Viral Spreading of Narratives

- Why / how do different arguments spread?
  - Key tradeoff: reputation vs. influence concerns
  - Social multipliers
- Negative narratives: giving an excuse $\Rightarrow$
  - Protects reputation of agent who behaves immorally $(a = 0)$
  - Impacts negatively audience's behavior, if it confronts similar choice
- Positive narratives or duties have the opposite two effects
- Therefore, expect that
  - When reputation (influence) concerns dominate, negative (positive) narratives are more likely to be shared than the opposite kind
- Examine under what circumstances each case obtains, both across individuals and as function of social structure
  - Abstract here from endogenous search: focus on strategic communication

Signaling and communication on a simple network

- Agents $i \in \mathbb{Z}$ on a line. Each may learn the (single) narrative $e \sim F(e)$ exogenously, with probability x (i.i.d.) –say, current media story
- Individuals can be "passive" (P) or "active" (A), and have values v_H or v_L
- Active agents choose $a_i = 0$ or 1, with action observed by successor $i+1$
  - Agent i cares about his reputation vis-à-vis agent $i+1$
  - Passive agents / principals: no reputation-relevant action



- Whether active or passive, if i has narrative e, can pass it on to $i+1$
  - Takes into account reputational impact + externalities this may induce through the moral/immoral actions of his successors $i+k$, $k = 1, \ldots +\infty$

- Active agents choose both what to do & say, passive ones only what to say
  - Active: $U = (v_i e - c/\beta)\, a + \mu \hat{v}(a,d) + v_i e N^i_{+/-}\ (a,d)$, $i = H, L$
  - Passive: $U = v_i e N^i_{+/-}(d)$, $i = H, L$
  - Action $a \in \{0,1\}$, disclosure $d \in \{e, \varnothing\}$ when knows e
- Don't know successor's A/P type: Markov process with serial correlation λ

$$\Pr[i+1 \in A \mid i \in A] = \Pr[i+1 \in P \mid i \in P] = \lambda,$$

  and independent probabilities $(\rho, 1-\rho)$ of $v_{i+1} = v_H$ or v_L
  - Steady-state: 50% A/P, with λ = degree of segregation, clustering
  - Extent to which i and $i+1$ engage in similar activities and have similar reputational concerns
  - Tree network: λ is % of i's direct successors who are of same type
- For simplicity, just two narratives:

$$e = \begin{cases} e_- & : \text{ prob } f_- \\ e_+ & : \text{ prob } f_+ \end{cases}, \quad \text{with } e_- < e^* < e_+ , \quad E_F(e) = e_0$$

Example 1: norms of gender relations in the workplace

- Men take actions or say things that affect the welfare of women (e). Are a priori uncertain (might say: "have no clue") whether those will be experienced as innocuous flirting, unwelcome advances, or 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 / people (v_H), others are indifferent or misogynistic (v_L), but all want to be seen as moral types

Example 2: ethnic majority/minority, redistribution

- Same framework applies to how a dominant national group will "treat," and justify treating, ethnic minorities or immigrants, or/and the poor.

- Would charitable contribution *a* or public transfer (tax compliance, vote):
  - Do significant good (e) for the poor: health, skills, children's education?
  - Be more likely captured by gvt / NGO bureaucracy, corrupt officials, wasted by recipients on drugs and alcohol, or collectively trap them into a toxic culture of welfare dependency?

Social transmission: basic tradeoffs and intuitions

- Passive agents have no reputation at stake, only influence concerns $\Rightarrow$ transmit positive narratives e_+, censor negative ones e_-
- Among active agents, both high and low types face potential tradeoff between reputation and influence concerns
  - The more moral v_H agents have stronger influence concern $\Rightarrow$ more inclined to spread positive narratives / refrain from spreading negative narratives
- Social multipliers: strength of influence concern also depends on how much further narrative is expected to be spread.
  - The more successors are, on average, likely to "serially" pass on excuses e_- for immoral behavior, the greater the social damage from sharing one $\Rightarrow$ each individual i will be less inclined to do so.

    $\Rightarrow$ Absolving, guilt-reducing disclosures should be strategic substitutes
  - The more successors are, on average, likely to "serially" pass on reasons e_+ for moral behavior, the greater the positive social impact of sharing one $\Rightarrow$ each individual i will be more inclined to do so.

    $\Rightarrow$ Responsibilizing, prosocial disclosures should be strategic complements

Analysis

- Focus on stationary equilibria such that:
  - Reputational motive prevails over influence one whenever they conflict: given an excuse e_-, both H and L active types invoke it and choose $a_i = 0$, even though may trigger chain of bad behavior from $i+1$ on. Most plausible case
  - In cases where H type does not know e (neither learned it by himself, nor heard from $i-1$), he chooses the same default action $a_H(\emptyset)$: the norm
- Distribution of beliefs:

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

- Virality factors or multipliers: N_-^i, N_+^i, for $j = A, P$: expected number of subsequent actions by $i+k$'s that will switch if j communicates e_- or e_+ to $i+1$. Pivotal for (say) N_-^i decisions $\Longrightarrow$ values at $v_H N^i e_-$ or $v_H \times N^i e_-$

Morality as the default behavior: it goes without saying

- Equilibrium where $a_H(\emptyset) = 1$: high types behave prosocially unless they learn of an exculpatory narrative e. Conversely, observing $a_{i-1} = 1$ reveals that they did not receive one

- When they learn e_- they choose $a_i = 0$ and pass it on, as do low types. Responsibilizing narratives e_+ are passed on by neither $\Rightarrow$

$$x_-^P = x + (1-x)(1-\lambda)x_-^A, \quad x_-^A = x + (1-x)\lambda x_-^A,$$
$$x_+^P = x, \quad x_+^A \equiv x.$$

- Study next the inferences that agents make when their predecessor chooses $a = 0$ without offering any rationale:
  - Posterior on his type: $\hat{v}_{ND} = v_L$
  - Posterior on social impact: $\hat{e}_{ND} > e_0$ if $i \in A$, $\tilde{e}_{ND} < e_0$ if $i \in P$

Proposition (morality as the default behavior)

In an eqbm where default (uninformed) action of high types is $a_H(\emptyset) = 1$:

1. *Positive narratives or duties,* e_+, *are transmitted by no one, since they do not change behavior* $(N_+^A = N_+^P = 0)$. *"It goes without saying."*
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 virality factor is*

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

   Such disclosures are therefore strategic substitutes.

4. *A greater degree of mixing between active and passive agents (lower* λ*) raises the aggregate provision of public good or externality*

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

- P agents act as "firewalls" for negative narratives, e_- : $x_-^A = x + (1-x)\lambda x_-^A$

Selfishness as the default behavior: silence is complicity

- Equilibrium with $a_H(\emptyset) = 0$: high types behave socially only in the presence of a responsibilizing narrative e_+, which they then pass on. Low types always withhold such narratives. Every active type still shares excuses e_- $\Rightarrow$

$$x_-^P,\ x_-^A : \text{unchanged from previous case}$$

$$x_+^P = x + (1-x)\left[\lambda x_+^P + (1-\lambda)\rho x_+^A\right],$$

$$x_+^A \equiv x + (1-x)\left[(1-\lambda)x_+^P + \lambda\rho x_+^A\right],$$

  - x_+^P and x_+^A : if $i-1 \in A$ and knows e_+, only high type will tell i

- "Influence multipliers" from sharing a narrative are now $N_-^A = N_-^P = 0$ and

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

$$N_+^A = (1-x)\left[\lambda\rho(1+N_+^A) + (1-\lambda)N_+^P\right].$$

- No-disclosure inferences: now $v_L < \tilde{v}_{ND} < \bar{v}$ on i's type, different $\hat{e}_{ND} > e_0$ on externality if $i \in A$, same $\tilde{e}_{ND} < e_0$ if $i \in P$

Proposition (selfishness as the default behavior)

In an eqbm. in which default (uninformed) action of high types is $a_H(\emptyset) = 0$ *:*

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 duties* e_+ *are transmitted by both passive agents and high-morality active ones. "Standing out for what is right"*
3. *The social impact of sharing a responsibility narrative is* $e_+ N_+^A$ *for an active agent and* $e_+ N_+^P$ *for a passive one, with virality factors* (N_+^A, N_+^P) *given recursively above. Such disclosures are therefore strategic complements.*
4. *A greater degree of mixing between active and passive agents (lower* λ*) raises the aggregate provision of public good or externality,*

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

- P agents now also act as "relays' for positive narratives, e_+, as do high-morality A agents

Main results on network structure

Proposition (average morality)

In either type of eqbm, more mixed interactions (lower λ) raise prosocial behavior.

- Agents whose actions and/or morality are not "in question" (irrelevant or unobservable) have no need for excuses $\Rightarrow$act as: (i) "firewalls" limiting diffusion of exonerating narratives; (ii) "relays" for responsibilizing ones; encourages high-morality actors to do so

Main results on network structure

- Recall that x_-^P, x_-^A are the steady-state probabilities with which any active or a passive agent will hear, directly or indirectly, about a negative narrative e_-, when there is one. For a positive one, they are x_+^P, x_+^A
- Differential exposure of A and P agents to each type of narrative, resulting in divergent views of what is (im)moral

Proposition (polarization of narratives and beliefs)

In either type of equilibrium, the information gaps between active and passive agents' awareness of narratives, measured respectively by $|\ln(x_-^P/x_-^A)|$ *for negative ones and* $\ln(x_+^P/x_+^A)$ *for positive ones, are both U-shaped in the degree of network segregation* λ*, with minimum of* 0 *at* $\lambda = 1/2$ *and maximum at* $\lambda = 1$.

- When A's and P's –e.g., men and women– interact mostly within segregated clusters $\Rightarrow$
  - Very different types of narratives will circulate between groups: mutual stereotyping
  - Men mostly sharing rationalizations for their behavior, which will be worse on average than under integration, and women mostly judging it as inexcusable.

4. Moral Standards: What Is Justifiable?

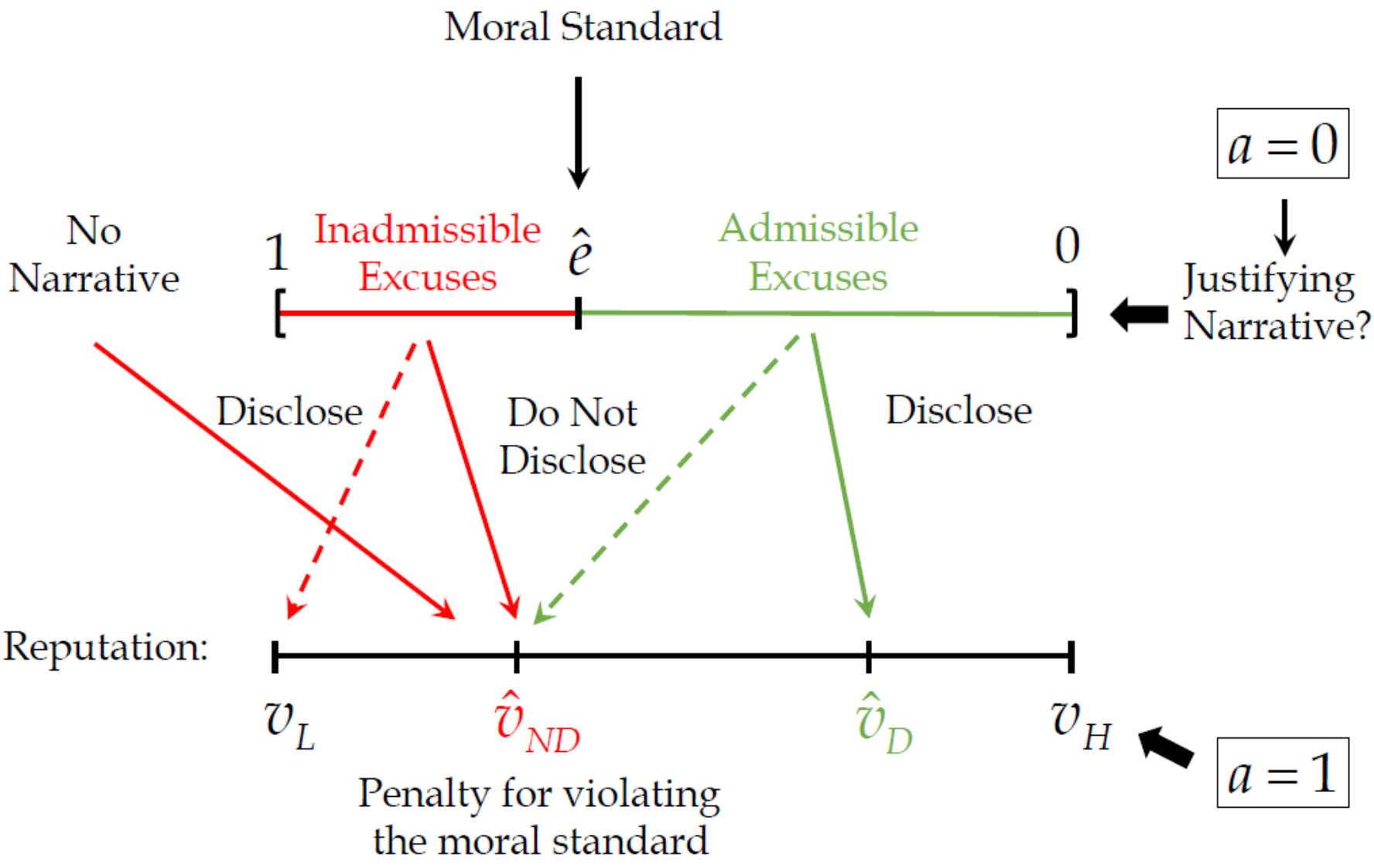
- Consider now case where signal or rationale about e arises from individual's own search for reasons to act morally, or not
- Enriched reputation channel: having an excuse means may have looked for one, says something about you. How strong does it then need to be?
  - Abstract from influence motive: single dyad of actor and "passive" audience
- Prior to acting, agent can learn e with probability x, at cost $\psi(x)$
  - ψ increasing, convex, Inada conditions
  - With probability $1-x$, learns nothing, denoted as $\emptyset \Rightarrow$ stays at e_0
- If has learned e, can disclose to audience (at ε cost),
  - Can also disclose after acting, but must credibly have obtained e before.
- "Default action" $\equiv a_H(\emptyset)$: based on prior only. Also refer to as "the norm"

Moral standards: basic questions

1. What is the norm / default behavior that will prevail?
2. How strong must an excuse in order to be "acceptable"?
3. How much stigma is incurred for failing to produce one?
4. Is producing an acceptable justification good or bad news about morality?
   - Better than choosing $a = 0$ without providing a justification
   - But likelihood of coming up with one will differ between high and low types: "Interesting how he/she always has a good excuse not to help!"

- Two key forces:
  - H type has decision-making incentives to search, plus possibly reputational one
  - L type always looking for excuses only, to pool with H when $a_H = 0$
- Equilibrium $\Rightarrow$ threshold strategies:
  - Both types will disclose and pick $a = 0$, when learn $e \leq \hat{e}$
  - Not disclose otherwise: $a_H = 1$ suffices to separate high type
- Moral standard $\hat{e}$ will be endogenous, differ from e^*
  - What factors make it lenient or tough?

Equilibrium moral standards



Straight arrows describe equilbrium play, dashed ones off-path deviations

Narratives and moral standards: main results

- Key factors determining whether a prosocial or antisocial culture prevails, high or low moral standards / tolerance for excuses :

1. Prior mean e_0 about whether actions have important or minor externalities:
   - Low $e_0 \longmapsto$ "selfish norm": default is $a_H(\emptyset) = 0$, need "reasons to act"
   - High $e_0 \longmapsto$ "prosocial norm": default $a_H(\emptyset) = 1$, need "reasons not to act"
   - Intermediate $e_0 \longmapsto$ coexistence of both norms, cultures
2. Tail risks (or option values) in uncertainty surrounding that question $\Rightarrow \hat{e}$

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

   - Example: suppose 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" $\Rightarrow$ even when such scrutiny reveals only far less serious concerns (isolated anecdotes, lowering e only slightly from e^*), these can become socially acceptable reasons for treating that group badly
   - There are now "excuses for having excuses," even weak ones $(\hat{e} > e^*)$

- $\mathcal{M}^-(\hat{e})$ low $\Rightarrow$ so are moral standards; $\mathcal{M}^-(\hat{e})$ or $\mathcal{M}^+(\hat{e})$ high, reverse

Proposition (**prosocial norm**)

For any prior e_0 high enough, 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 ($\hat{v}_{ND} = v_L$)

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

$$e^* - \mathcal{M}^-(e^*) > \mu\rho(v_H - \bar{v})/v_H, \tag{1}$$

   (a) The high type is more likely to search for narratives: $x_H > x_L$, and so producing one improves reputation, $\hat{v}_D > \bar{v}$

   (b) The option value of finding potentially strong reasons for not taking the moral action makes coming up with even a relatively weak one less suspect $\Rightarrow$ lower moral standard: $\hat{e} > e^*$.

2. *If $F(e)$ is sufficiently bottom-light, reversing (1):*

   (a) The low type is more likely to search for narratives: $x_H < x_L$, and so producing one worsens reputation, $\hat{v}_D < \bar{v}$

   (b) Because most excuses one could hope to find are weak ones, coming up with even a strong one is suspicious $\Rightarrow$ higher moral standard: $\hat{e} < e^*$.

Proposition (**selfish norm**)

For any e_0 low enough, there exists an equilibrium where abstaining is the default (uninformed) choice of the high type 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 (especially when F is top-heavy), and as a result lowers the moral standard ($\hat{e} > e^*$).*

Proposition (multiple norms and meanings of excuses)

Let $\psi'(1) = +\infty$. There is a range $[\underline{e}_0, \bar{e}_0]$ such that, for any prior e_0 in that interval, a prosocial and a selfish norm coexist.

5. Narratives Versus Imperatives

- Modeled narratives as persuasive *per se:* arguments about a decision that act like hard information -though may or may not be
- Imperatives as polar opposite: broad recommendations, without giving reasons: "just do it", "thou shalt not kill". Soft information / pure cheap talk
  - Now a dual influence channel. Focus on single principal-agent dyad (P/A).
- Costs of imperatives:
  - May be less effective, if lack of congruence between principal and agent: impact hinges critically on who issues them. Need moral legitimacy
  - More rigid (non contingent) $\Rightarrow$ will sometimes give the wrong decision
- Benefits of imperatives:
  - Less fragile to misunderstanding, interpretation uncertainty, risk of debunking
  - Allow pooling of states in which agent would be reluctant to act with others in which he would be willing

    $\Rightarrow$ Provision of externalities over a broader range of e's

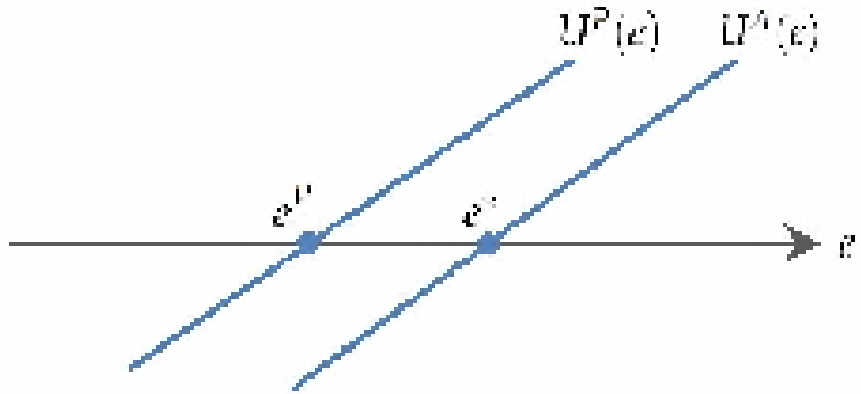
Preferences and information

- Principal: society, religious leaders, parents, ex-ante self. Wants to promote creation of positive externalities / internalities, discourage negative ones.
- Example:

$$W = [we + (\bar{v}e - c)] \times a,$$

  - $w = 0$: parents maximizing child's welfare; $w = 1$: utilitarian social planner
  - $w = \pm\infty$: moral / immoral entrepreneur with no empathy for agent

- More generally, principal's utility $aU^P(e)$, H agent's utility $aU^A(e)$



- Principal has / finds signal or narrative: $e \sim F(e)$ on $[0, 1]$, mean e_0
- Assume $e_0 < e^P < e^* \Longrightarrow$ inducing agent to choose $a = 1$ will require a convincing argument. Share narrative / reasons, or issue imperative?

Coarse versus noisy communication

- Principal can disclose narrative e (if has $e > e^*$), or/and issue imperative: "do $a = 1$." Agent chooses $a \in \{0, 1\}$.

- Narrative: usable by anyone who has it, but may be misunderstood / fail to convince:
  - Agent $\begin{cases} \text{understands argument with probability } \xi \Rightarrow \text{ posterior } e \\ \text{fails to understand } (\emptyset) \text{ with probability } 1-\xi \text{ (can be } \simeq 0). \end{cases}$

- Imperative: cheap talk, so credibility constraint. Equilibrium requires:

  (IC) Anticipating obedience, principal recommends $a = 1$ iff

$$U^P(e) \geq 0 \quad \Leftrightarrow \quad e \geq e^P$$

  (CR) Obedience: agent picks $a = 1$ when told to pick $a = 1$:

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

  - Again, tail moment of $\mathcal{M}$ is critical ("top/bottom heaviness" results)

Proposition (clarity vs. credibility)

Suppose that any narrative involves at least a slight probability of miscommunication. There is a unique equilibrium:

1. *If $\mathcal{M}^+(e^P) \geq e^*$, the principal issues an imperative whenever $e \geq e^P$, and does not communicate otherwise.*

   If $\mathcal{M}^+(e^P) < e^*$, she discloses narrative whenever $e \geq e^*$, does not communicate otherwise. Achieves $U^P(e)$.

2. *The use of imperatives (vs. narratives) 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 (higher e^P)*
   - *She is better informed about externalities from their actions ($F(\cdot)$ more informative / top heavy), and not too pessimistic*
   - *These externalities are likely to be important a priori (uniform shift in $F(\cdot)$, satisfying MHR).*

Congruence and flexibility

- When given narrative e, the agent may do his own additional thinking, search for counter-argument, etc. $\Rightarrow$ arrives at final assessment of the externality $\epsilon \sim H(\epsilon|e)$, with $E(\epsilon|e) = e$ and $H(e^*|e) < 1$ for all e, such that:
  - Increase in e shifts distribution of ϵ to the right, in the sense of MLRP
  - ϵ is a sufficient statistic for $(\epsilon, e) \Rightarrow$ principal also wants to evaluate final payoffs, and thus agent's choices, according to ϵ
- Alternatively, may prefer to issue imperative to "$a = 1$" over some subset of states, denoted I. Equilibrium requires:

$$\text{(CR):} \quad E[e|e \in I] \geq e^*,$$

$$\text{(NVA):} \quad \int_0^{e^*} U^P(\epsilon) dH(\epsilon|e) \geq 0 \quad \text{for all } e \in I,$$

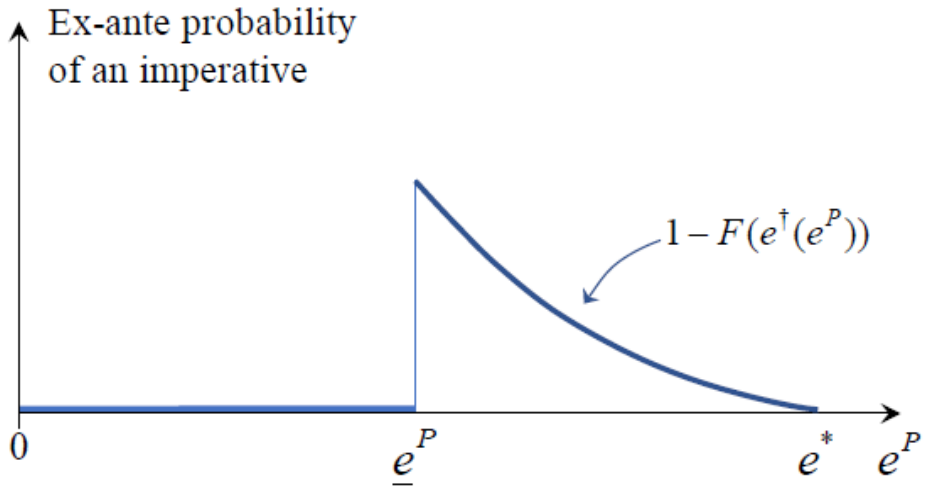
  - Advantage of narrative strategy is flexibility, valued by both parties: when $e > e^P$ but agent's additional information leads to posterior $\epsilon < e^P$, rightly chooses $a = 0$, whereas under (effective) imperative would have chosen $a = 1$

Proposition (congruence and flexibility)

Suppose the agent can use private information to refine the principal's narrative, $e \longrightarrow H(\epsilon|e)$, so imperatives are costly in terms of flexibility. Define $e^{\dagger} > e^{P}$ by

$$\textit{Option value of "independent thinking"} \equiv \int_0^{e^*} U^P(\epsilon) dH(\epsilon|e^{\dagger}) = 0.$$

1. *Imperatives are used in equilibrium if and only if $\mathcal{M}^{+}(e^{\dagger}) \geq e^{*}$. Imperative is then issued when $e \geq e^{\dagger}$, while for $e < e^{\dagger}$ narrative is disclosed*
2. *The probability of an imperative being issued is hump-shaped in congruence (e^{P}). Applies in particular to self control.*



Directions for future research I

- 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.
  - Taking as a primitive some class of arguments that "work" (a least probabilistically) 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?
  - Important elements of both heuristic and motivated or wishful thinking. Models of these exist, could be combined with present one
  - But other aspects, still to be better understood / studied
- Competing narratives: parties with conflicting interests, opportunistic "narrative entrepreneurs," will offer very different rationales for right/wrong
  - What factors then make one story more compelling than the other?

Directions for future research II

  - Related: politics (fake news, focus of the debate), construction of identity, in-group/out-group conflict, design of religious and ideological doctrines (complementarity between narratives and imperatives)

- Preference disagreements over what constitutes a moral act / negative vs positive externality, even under full information (or: heterogenous priors) $\Longrightarrow$ social image becomes multidimensional and subgroup-dependent
  - Critical for "hot" societal issues such as abortion, gay rights, gun control, etc.
  - Model captures a good part of this already (find ways of reducing to a single-dimensional problem), but not with full generality
  - Endogenous sorting a key issue

- Experiments...