SAFE SPACES: SHELTERS OR TRIBES?

Jean Tirole

NBER's SI 2021 IT and Digitization Workshop

July 22, 2021

[This is a long version of the slides. An abridged version will be presented on July 22.]

Work in progress! Comments particularly welcome

Toulouse School of Economics

IAST 10 years of interdisciplinary research

I. INTRODUCTION

Research is about the allocation of our life between private and public spaces.

This allocation reflects

- *Technological evolution* (AI, facial recognition, smart phones, social networks…) $\Rightarrow$ expansion of the public sphere.

  Not a random one: The selective relationships of our private sphere are (endogenously) biased towards like-minded individuals.

- *Laws* (EU: 2014 ECJ decision on right to oblivion, 2016 GDPR, 2021 AI Act) and *norms* (doxing, outing, paparazzi).

- *Individual choices*: two behavioral reactions (retreat in safe space, change in behavior).

Consensual issues

Much of the theoretical and empirical attention has been focused on broadly consensual behaviors

- Agreement on what is right or wrong (pollution, crime; charitable contributions, public good provision, voting, blood donation…)

Divisive issues (society and epoch specific)

- Politics
- Sexual orientation
- Religion, secularism
- Vegans and meat-eaters, abortion, social roles, corrida/boxing, religious slaughtering of an animal, vaccines

Image/self-presentation concerns differ!

Description of image concerns: consensual behaviors

Agent i takes action a_i, has privately known type v_i on $\mathbb{R}^+$ (e.g., extent of prosociality/other-regarding preferences) drawn from cdf $F(\cdot)$.

Reputational payoff depends on posterior beliefs $F(v_i|a_i)$; often summarized by representative type $\hat{v}_i \equiv E[v_i|a_i]$.

- Image payoffs: pure image concerns or functional (matching opportunities, reciprocity, etc.)
- In some papers, agent i can affect the visibility of her action to her potential audience.
- Accommodates both demand for a high reputation (say wants $\hat{v}_i$ as high as possible, as in Bénabou-Tirole 2006) or demand for an intermediate reputation (Bernheim's 1994 theory of conformism).

Description of image concerns: divisive behaviors

Agent i takes action a_i, has (horizontal) type $v_i \in \mathbb{R}$. Again, wants to ingratiate themselves with audience(s).

Two new features

- *Receiver-contingent judgment.* Reputation is in the eyes of the beholder. The same behavior is frowned upon by some, liked by others.
- *Differential disclosure.* Whether i's behavior is observed will depend on type v_j of receiver (even though v_j is not directly observed by i).

Example of formalism: i's reputational payoff with j

$$r(\hat{v}_{ji}, v_j)$$

i's type as perceived by j: differential disclosure

in the eyes of the beholder

Preview of equilibrium behavior

Demand for selective disclosure (for safe spaces): Image concerns imply that we would want our behavior to be known

- to the in-group of like-minded individuals choosing the same behavior
- not to out-group: full transparency may make us shy to act

Retreat in a safe space, physical (home, private club, church, masonic lodge, bullfight ring, political party…) or virtual (Facebook group) generates less hostility: *shelter aspect*.

But it comes with private costs

- deviation of behavior from authenticity
- hiding costs
  - reduced use of public space (exogenous hiding cost)
  - forgoing desirable relationships and diversity of social graph (endogenous hiding cost).

Welfare impact of technology and laws: Does laissez-faire generate too little or too much transparency?

Other considerations:

(1) Social benefits of safe spaces on image side
   - Pure reputation stealing ("positional image"), in which case welfare effect only through impact on behavior
   - Or reduce DWL (ostracism/discrimination/hatred fueling/violence…by employers/coworkers, anonymous hatemongers, blackmailers, indelicate governments)

(2) Collateral social costs: once in a safe space, one-upmanship/holier than thou attitude one-sided narratives, hate speech, conspiracy theories, Facebook groups
   - *Tribal behavior:* voluntary or enforced by threat of outing/exclusion.

Relationship to the literature

Very large theoretical and empirical literature on prosocial behavior

- Prediction that giving a socially-valued behavior more visibility makes it more prevalent [Ali-Bénabou 2020]. Conversely, reduces occurrence of behaviors that are frowned upon [Daughety-Reinganum 2010 on refraining to check in rehab center or disclosing information about health; Jann-Schottmüller 2020 on chilling effect]
- Strong evidence on impact of visibility
- Won't be true for a divisive behavior

Literature on conformity [Bernheim 1994, Manski-Mayshar 2003, Kuran-Sandholm 2008, Michaeli-Spiro 2015, 2017]

Literature on countervailing incentives [Gertner et al 1988, Spiegel-Spulber 1997, Austen-Smith-Fryer 2005, Bar-Isaac-Deb 2014, Bursztyn et al 2017, Bouvard-Levy 2017]

Rather different modeling, questions and conclusions here.

Broader social-science debate on which of privacy and transparency best promotes social welfare

- Philosophers' positive connotation of authenticity: associated with emancipation brought about by privacy, a view that has much influence on current laws and privacy activism.
- Sartre. Williams: "To act morally is to act autonomously, not as the result of social pressure".

II. DIVISIVE BEHAVIORS

Actions

Mass 1 of agents
Agent i takes action $a_i \in \{-1, 0, +1\}$

- $a_i = 0$: passive/stay neutral
- $|a_i| = 1$: acts, at cost $c \geq 0$ (time, cost of donating to activity, demonstration,...)

Non-image payoff

$$v_i a_i - c|a_i|$$

Preference heterogeneity

Type v_i private information, drawn from $F(v_i)$ on $\mathbb{R}$. Cumulative distribution is unimodal and symmetric around 0; has a mean (necessarily 0).

Image concerns

- Reputational payoff vis-à-vis j: $r(\hat{v}_{ji}, v_j)$ where $\hat{v}_{ji} = E_j[v_i]$ (dependence on j reflects j's information about a_i).

  [Later on, alternative formulation: reputation as a random, rather than representative member of perceived group. Then, i's reputational payoff with j is $\int_{-\infty}^{+\infty} r(v_i, v_j) dF_j(v_i)$.]

- Agent i's overall reputation payoff in society

$$R_i \equiv \int_{-\infty}^{+\infty} r(\hat{v}_{ji}, v_j) dF(v_j).$$

Payoffs

Self-presentation/hiding cost h_i (see later). Agent i's utility

$$u_i = v_i a_i - c|a_i| + R_i - h_i.$$

Equilibria

Symmetric equilibrium. For some $v^* \geq 0$

$$a_i = \begin{cases} 1 & \text{for} \quad v_i > v^* \\ 0 & \text{for} \quad -v^* < v_i < v^* \\ -1 & \text{for} \quad v_i < -v^* \end{cases} .$$

Disclosure behavior will be symmetric as well.

[There will exist no asymmetric equilibrium.]

Assumptions on image concerns

Assumption 1 *(symmetry).*

For all $(\hat{v}, v)$,

$$r(-\hat{v}, -v) = r(\hat{v}, v).$$

Assumption 2 *(distaste for dissonance).*

Ceteris paribus, agents want to ingratiate themselves with others. Suppose that $v > 0$. *Then for all* $\hat{v} < v$

$$r_1(\hat{v}, v) > 0.$$

Assumption 3 *(concavity).*

Perceived ideological differences have an increasing marginal cost: for all $(\hat{v}, v)$,

$$r_{11}(\hat{v}, v) \leq 0.$$

Assumption 4 *(benefit from being perceived by the in-group as representative of the in-group rather than as the average type in the population).*

Let $M^+(v^*) \equiv E[v|v \geq v^*]$. *An agent picking* $|a_i| = 1$ *gains from being perceived by her in-group as the mean type of the group rather than as the average type in the population: for all* $v^* \geq 0$,

$$\int_{v^*}^{+\infty} [r(M^+(v^*), v) - r(0, v)] dF(v) > 0.$$

Examples satisfying 4 assumptions

(1) *Positional image*

$$r(\hat{v}, v) \equiv \mu\theta(v)\hat{v}$$

when $\theta(\cdot)$ antisymmetric (with $\theta(0) = 0$) and increasing. So $r_{11} = 0$. Image is constant-sum in society (only reputation stealing).

(2) *Placating image concerns*

Want to be perceived as close in values as possible to audience:

$$r(\hat{v}, v) \equiv -\mu(|\hat{v} - v|)^p$$

for $p \geq 1$. Modified L^p norm.

Alternatively, one can define total reputational payoff directly (non-additivity)

(3) *True L^p norm*

$$R_i \equiv -\mu\Big(\int_{-\infty}^{+\infty} |\hat{v}_{ji} - v_j|^p \, dF(v_j)\Big)^{1/p}$$

(4) *Maximum norm*

F have finite support $([-V, V])$

$$R_i \equiv -\mu \max_{v_j} |\hat{v}_{ji} - v_j|.$$

Focuses on most hostile.

Upper bound on welfare W

Under these assumptions, full privacy yields the highest possible welfare

- Authenticity $v^* = v^{fp} = c$
- Total agent reputational payoff is maximized.

However full privacy is not an equilibrium: The highest privacy level will correspond to a safe space equilibrium, with second-best total agent reputational payoff, but low authenticity (plus collateral damages).

Demand for reputation

Thought experiment: suppose that action $a_i \in \{-1, +1\}$ (chosen by $|v_i| \geq v^*$)

- is observed by peers/in-group $\equiv$ those who pick the same action
- is hidden from outgroup with probability x.

Proposition *(demand for joining a safe space)*

Under Assumptions 1 through 4, and ignoring any cost of self-presentation, an agent i who selects $|a_i| = 1$ strictly prefers to disclose her behavior to her peers, and prefers not to disclose her behavior to non-peers (strictly so unless $v^* = 0$ and $x = 0$); and so $x_i = 1$.

III. THE EMERGENCE OF SAFE SPACES AND THEIR IMPLICATIONS

Costless self presentation $(h = 0)$

Previous result $\Rightarrow$ $x = 1$ is an equilibrium.

Equilibrium cutoff $v^* = v^s$ $(= 0$ when c is sufficiently small). When strictly positive:

$$v^s - c + \int_{v^s}^{+\infty} [r(M^+(v^s), v) - r(M^-(v^s), v)] dF(v) = 0$$

where $M^+(v^s) \equiv E[v|v \geq v^s]$ and $M^-(v^s) \equiv E[v|v < v^s]$

Implies that $v^s < c$.

- Comparison with two polar benchmarks:

  Full privacy (hypothetical): $v^{fp} = c$ (authenticity)

  Transparency (will occur for high hiding costs)

$$v^t - c + \int_{-\infty}^{+\infty} [r(M^+(v^t), v) - r(0, v)] dF(v) = 0$$

  so

$$v^t \geq c; \quad \text{(strictly so when } r_{11} < 0\text{)}$$

- *Social pressure externality (amalgam effect) under safe spaces*

  Passive agents receive lower payoff than under full privacy or transparency: they are viewed suspiciously by both sides.

Costly self-presentation

Hiding from out-group costs $h \geq 0$
Exogenous cost for now (= not using the public space)
Cutoff's net benefit from acting in safe space

$$S(v^*, x) \equiv v^* - c + \underbrace{R_1^s(v^*, x)}_{\substack{\text{total reputation} \\ \text{when } a_i = +1 \\ \text{and safe space}}} - \underbrace{R_0(v^*, x)}_{\substack{\text{total reputation} \\ \text{from} \\ a_i = 0}}$$

Cutoff's net benefit from acting transparently

$$T(v^*, x) \equiv v^* - c + \underbrace{R_1^t(v^*)}_{\substack{\text{total reputation} \\ \text{from } a_i = +1 \\ \text{transparently} \\ \text{(does not depend on } x\text{)}}} - R_0(v^*, x)$$

Safe space equilibrium $(x = 1)$ satisfies

$$S(v^s, 1) - h = 0 \geq T(v^s, 1)$$

Transparency equilibrium $(x = 0)$ satisfies

$$T(v^s, 0) = 0 \geq S(v^s, 0) - h$$

Mixed equilibrium $(0 < x < 1)$ satisfies

$$S(v^m, x) - h = T(v^m, x) = 0.$$

Assumption 5

$S(v^*, x)$ and $T(v^*, x)$ *are strictly increasing in* v^* *for all* x.

Ensures uniqueness, satisfied if image concerns (μ) not too large and either (a) finite support or (b) $f(v)v^p$ bounded for true L^p norm (no fat tails).

Assumption 6

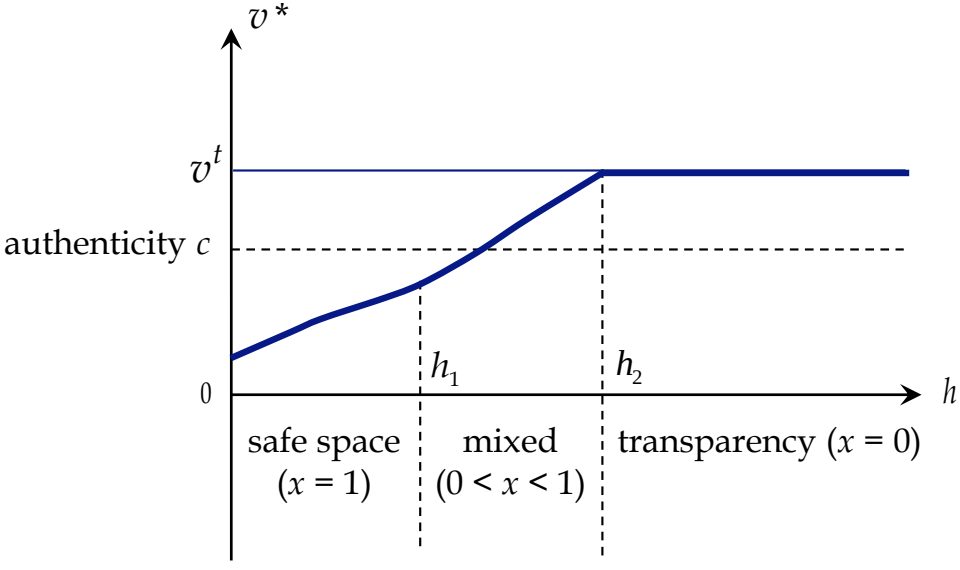
$c > R_1^s(0,1) - R_0(0,1)$.

Only to shorten exposition (avoids corner solution $v^s = 0$).

Proposition

Unique equilibrium; is symmetric. Characterized as in Figure below.

General case

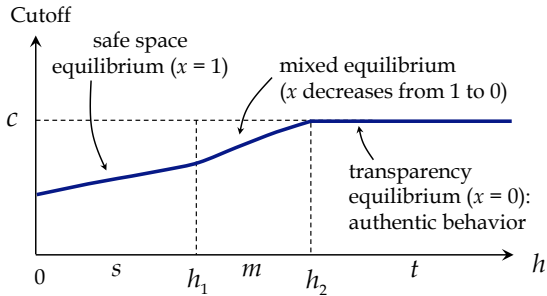


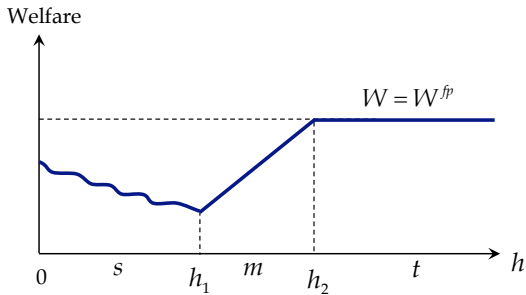
Positional image

$$r(\hat{v}, v) = \mu\theta(v)\hat{v}$$

Transparency maximizes welfare:

- authentic behavior ($v^t = c$ as $\int_{-\infty}^{+\infty} \mu\theta(v)\hat{v}dF(v) = 0$ for all $\hat{v}$)
- image is positional (zero-sum game)

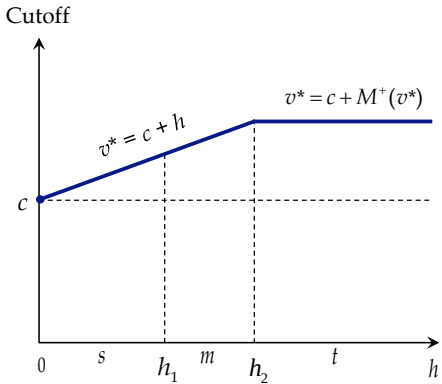


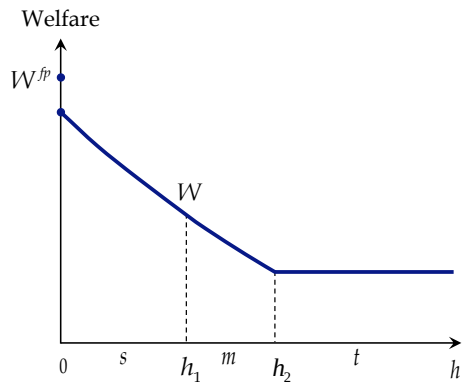


Maximum norm

(1) Level of activity always lower than the authentic level: $v^* \geq c$

(2) Welfare continuously decreasing in h. Making it more difficult to hide forces socially undesirable transparency.





Dynamics of divisive behaviors

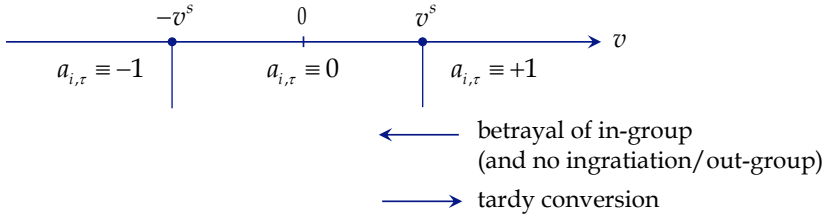
Repeated game $\tau = 0, 1, \ldots, +\infty$

Sequence of actions $a_{i,0}, a_{i,1}, \ldots \in \{-1, 0, +1\}$

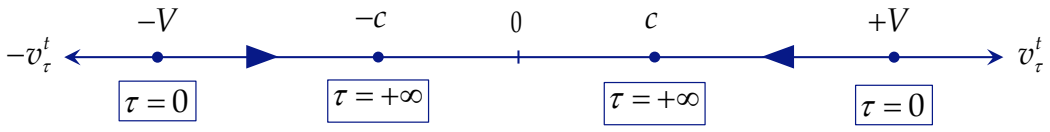
Look at polar cases: h low (safe spaces) and h high (transparency).

Payoff $\sum_{\tau=0}^{+\infty} \delta^{\tau}[v_i a_{i,\tau} - c|a_{i,\tau}| + R_{i,\tau}]$

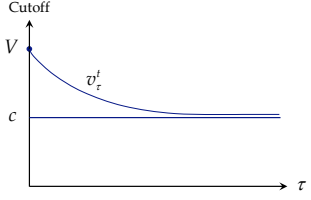
Low hiding costs: stationary outcome = repeated static outcome $(v_{\tau}^{s} \equiv v^{s})$



High hiding costs: Coasian dynamics: v^t decreases over time (more and more pressure to act over time). Example: continuous time, max norm



that is:



Once agent has shown "not to be an extremist", she can behave more authentically.

Reputation as a random member of a group

Reputational payoff

$$\int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} r(\tilde{v}, v_j) dF(\tilde{v}|v_j) \right] dF(v_j)$$

- Same if r linear in $\hat{v}$ (i.e. $r = \mu\theta(v)\hat{v}$)
- Positional image (constant sum) more generally (i.e. even though $r_{11} < 0$)
- Characterization the same as the previous one for the positional image.

IV. EXTENSIONS, APPLICATIONS AND DISCUSSION

(1) Endogenous social graphs

Assumption (too extreme): creation of a safe space requires social graph that is composed solely of like-minded agent $\Rightarrow$ must morph social graph to a more homogeneous one

- Paper argues that a good representation of the cost of moving from graph f to graph g is (proportional to) the L^1 distance:

$$\|f - g\| \equiv \int_{-\infty}^{+\infty} |f(v) - g(v)| dv.$$

May come from either loss of diversity or mere cost of changing friends.

Two new features:

- Strategic complementarities
- Lock-in if cost of changing friends is one-shot rather than recurrent.

(2) Outing and coming out

Outing (being kicked out of safe space): most often of a celebrity. Clear cost, but where is the demand for outing?

- Conjecture: makes the community more mainstream, less threatening.
- Outings may then trigger coming outs.

(3) Collateral damages: from shelter to tribe

- Add an additional action/signal (spreading -or refraining from spreading- narratives, engaging in hostile action against out-group…)
- Once in safe space
  - strong incentive for one-upmanship (voluntary signaling)
  - vulnerable to pressure from in-group or its sponsor: threat of exclusion or outing.

IV. SUMMARY

Platforms and governments increasingly trespass on our privacy.

- The public policy debate emphasizes the benefits from *privacy*: It allows us to behave authentically, without fear of hostility from non-liked-minded fellows.
- Much economics literature emphasizes the benefits of *transparency*: It makes citizens, workers, suppliers, and governments more accountable for their behavior.

This work studies divisive issues

- Politics, religion, sexual orientation, social roles, vaccines, abortion, corrida/boxing…

To that purpose, it develops a new framework for thinking about reputational concerns

- Opinions about an agent are contingent on the audience's views ("in the eyes of the beholder")
- Information about an agent is also contingent on audience's views (endogenously selective disclosure).

Insights

1. The proper comparison is often not between full privacy and transparency
   - Agents want to ingratiate themselves with their in-group, which they discover by joining a safe space.
2. The joining of a safe space captures the quest for a shelter as envisioned by the privacy advocates, but implies "reputation stealing" externalities.
3. Welfare implications depend on the concavity of the reputational payoff
   - When hiding in a safe space is mainly about stealing reputation from others (positional image), transparency is socially desirable, as it *reduces* posturing/promotes authenticity
   - When the reputational payoff r is more concave, safe spaces act as shelters against value destruction (discrimination, violence…) and socially dominate transparency.

4. Safe spaces cannot be assessed without considering their collateral damages. Members may engage in one-upmanship
   - either voluntarily, to prove that they are the true believers
   - or prompted by the safe-space gatekeeper or members threatening an outing or an exclusion.

   Either way, safe spaces are a threat for social cohesiveness and democracy.