Forthcoming, *Journal of Political Economy.*

Safe Spaces: Shelters or Tribes?[*]

Jean Tirole[†]

January 2, 2026

Abstract: The paper develops a framework for thinking about social behavior in the realm of divisive issues. When concerned that others may hold different views on what's right and wrong, we may change our behavior, or else join a safe space at the cost of a reduced use of public spaces and an insular social graph. This paper applies the framework to study the emergence of safe spaces of like-minded individuals, their dual role as shelters and tribes, and their societal consequences.

Keywords: Privacy, divisive issues, safe space, in- and out-groups, social graph, authenticity, polarization, outing.

JEL numbers: D64, D80, K38.

[*] The author acknowledges funding from the European Union (ERC, HWS, grant n° 101098319) and the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010. He gratefully acknowledges the financial support of the TSE Digital Center (the list of sponsors is available at https://www.tse-fr.eu/digital). He also thanks Roland Bénabou, Emir Kamenica, Karine Van der Straeten, Glen Weyl, two referees, and participants at various seminars and conferences for helpful comments, and to Bin Cheng, Paul-Henri Moisson and especially Amirreza Ahmadzadeh for excellent research assistance.

[†] Toulouse School of Economics (TSE) and Institute for Advanced Study in Toulouse (IAST).

1 Introduction

Numerous behaviors and opinions are approved by part of our audience and vilified by others. Our religious or political views, our sexual orientation or gender identity, our attitudes towards abortion, immigration, surrogate motherhood, transgender rights, DEI, GMOs, vaccines, or social roles are all matters of contention. How should we think about divisive issues? Conceptually, society's diversity of preferences is already captured in Hotelling-style models. But there is no arguing about tastes in such models: Those uninterested in stamp-collection just do not collect stamps; they do not stigmatize, marginalize, maltreat or discriminate against stamp collectors. In contrast, I will focus on two signatures of divisive issues, expressing the conflict between our social identity and the identity of the self.

First, the longing for "normality" makes us change behavior, privileging our social image over our preferences. We abandon our religious practices, conform with our expected social role, or have the unwanted child. Second, and alternatively, we keep the same behavior but hide it: We alter our use of the public space to avoid revealing our sexual orientation or our political opinions; we refrain from militantism or simply not dare to express our views in public; we incur costs of mascarading as someone "normal" (as in the case of an homosexual marrying heterosexually, or even endorsing anti-homosexuality laws).

While remaining authentic we may thus hide from the "out-group", defined as the individuals behaving differently than we do, by entering a "safe-space" of like-minded individuals (the "in-group"). The in-group engages in a common activity, whether virtual (a social-network group[1]) or real (a family house, a church, a club, a masonic lodge, a bullfight ring, a secret society, an ethnic neighborhood, or a political party, all occupied only by peers, like in gated communities or the rural or urban communities where some actors of the 1968 contest movement took refuge from an oppressive society). We may alternatively narrow our social graph to restrict information leakage, by foregoing potential friendships in the out-group (in-group acquaintances are more reliable as they experience more empathy for our preferences and furthermore can be outed in retaliation if they out us[2]) or turning to shadow suppliers (as when drug users or aborting women resort to costly and untrustworthy providers)[3].

[1] A virtual space of like-minded agents is a platform on which agents exchange with each other. An analogy with Facebook groups may be useful. Facebook recommends to its individual users groups of users who are like-minded. Indeed, a 2016 internal Facebook presentation on extremism in Germany commented on the platform's creation of communities around shared interests through its recommendation system and stated that "64% of all extremist group joins are due to our recommendation tools."

[2] As Simmel (1906) notes, "The first internal relation that is essential to a secret society is the reciprocal confidence of its members... Its elements may live in the most frequent commerce, but that they compose a society -a conspiracy, or a band of criminals, a religious conventicle, or an association for sexual extravagances –may remain essentially and permanently a secret."

[3] I are interested in actions for which there can be a meaningful safe space. Consider a pro-abortion stance. A safe space, in which this view can be expressed, may make sense in order to avoid being harassed by the other side. In contrast, when trying to inflect public policy by demonstrating, then joining a safe

A large theoretical and empirical literature focuses on "consensual environments", in which perceptions are uniform: Everyone agrees on what is virtuous and all actors wish to be perceived as virtuous. In contrast, little is known about "divisive-issue environments". Section 2 proposes a framework for thinking about the endogeneity of our private sphere in environments in which issues are divisive. Agents have types/identities that describe their preferences regarding some issue; I make reasonable, reduced-form assumptions regarding image payoffs. No model can fully capture the richness of all examples listed above, but two variants of a single paradigm go a long way to achieving this:

The *"trinary action"* version, which may capture politics, abortion, GMOs, DEI, vaccines, or social roles, stages two antagonist blocks (left/right, progressive/conservative, pro-choice/pro-life, anticlerical/deeply religious...), with moderates in between (which does not mean that they emerge unscathed, as we will see). There are two possible actions, advocating or working for a left-wing or right-wing viewpoint or policy, plus a moderate stance. So, there are two partisan or activist actions $a = -1$ and $a = +1$, as well as a neutral one $a = 0$.

In the "*binary action*" version, a majority favors action $a = 0$, and (rightly or wrongly) frowns upon the preferences of a minority. There are two relevant actions for the minority agents, who prefer $a = 1$: follow one's intrinsic preferences ($a = 1$) or comply with the injunctive norm ($a = 0$; the majority's choice between the two is a no-brainer). I here have in mind a sexual minority, a religious or agnostic minority in a nation with a state religion, an opposition to an autocratic regime, a minority of fans of violent activities or shows (boxing, corrida, MMA, guns)... Technically, the binary-action model is a special case of the trinary-action one, with a very asymmetric distribution of preferences.

I assume that an activist choosing action $|a| = 1$ can, at cost $h \geq 0$, conceal this action from the out-group (the agents who make a different choice). Furthermore, because partisans, when joining safe spaces, will be shown to exert negative image externalities on the moderates, an important assumption will be that the moderates cannot prove their moderateness. This assumption has two formally equivalent motivations:

- *Passivity.* One interpretation of the moderate choice is that it involves no action and exhibits plausible deniability (lack of time, other interests...): One cannot prove that one does not go to church or support a political organization. The activists' hiding cost h then may refer to a limited use of the public space or a narrow social graph confined to like-minded peers.

- *Duplicity.* The alternative interpretation is that the action $a = 0$ is instead an active one, that can be observed. The activists, i.e. those who choose $\mid a \mid= 1$, have two possible image-management strategies. Their choice can be the end of it, de facto revealing publicly the agent's action: The absence of a public choice of $a = 0$ betrays choice $\mid a \mid= 1$. Or else they can additionally engage in action $a = 0$ as well,

space would defeat the purpose, at least for the leaders of the movement (mere demonstrators may just wish that their anonymity will be preserved from hostile individuals).

and thereby incur some duplication cost $h > 0$ - such a strategy makes sense only if they consume action $\mid a \mid= 1$ secretly, i.e., in a safe space (transparency would defeat the purpose). For example, in the binary-action version, choosing action $a = 0$ on top of $a = 1$ may stand for marrying heterosexually for a homosexual or practicing one's true religion secretly while publicly practicing the dominant religion. Duplicity may also consist in expressing average-image-enhancing views in public while expressing different views in the privacy of a safe space. The cost h might then be a psychological discomfort or a time cost associated with the two-faced behavior.[4]

Regardless of the relevant version (trinary or binary), individuals' behaviors consist in the choice of behavior and, when acting, whether to incur a cost of concealing one's behavior from the out-group. I make assumptions that guarantee that (a) an agent is always eager to reveal to in-group members that "*he is one of them*" by disclosing his behavior to them and (b) the agent prefers not to antagonize members of the out-group by revealing his behavior, and so he does not disclose his behavior to them in the absence of a hiding cost.

Starting with the trinary model (the symmetric version of the overall one), my first result is that the equilibrium is unique and depends on the hiding cost in a natural way. Behavior involves acting in a safe space (i.e. covertly) for low hiding costs, acting transparently for high hiding costs, and playing a mixed disclosure strategy in between. Next, I compare behavior with the "authentic" one that would prevail if no information about behavior ever filtered out, even to the in-group[5] (a thought experiment that, as we saw, is not an equilibrium outcome). High hiding costs cause transparency and make agents scared to act. In contrast, low hiding costs embolden the agents to act (in safe spaces); they then exert negative externalities on moderates: An amalgam effect ("*He who is not with Me (once and for all on My side) is against Me*") compels moderates to take side when they would not want to.

Section 3 provides micro-foundations for the image payoffs posited in Section 2. The agent's overall image payoff is obtained by aggregating bilateral reputation payoffs: Agents have image payoffs with each other, that may for instance depend on the perceived distance between their views (say, according to the L^p norm). I make reasonable assumptions on bilateral reputation payoffs and check that the resulting overall image payoff satisfies the assumptions of Section 2. With some caveats, I then look at utilitarian welfare obtained by adding up agents' utilities and ask when the shelter benefits enjoyed by users of safe spaces dominate the amalgam losses incurred by moderates. When reputations are redistributive, i.e. have no aggregate image consequences (image is positional), the amalgam effect

[4] It could also correspond to the moral condemnation of the duplicity within the group of agents who are aware of it (namely the agent's peers in the safe space). Note however that these agents are themselves duplicitous and therefore reluctant to condemn the agent's deceit.

[5] Under my assumptions, social welfare is maximized in this fictitious "full-privacy case", for two reasons: (1) Behavior is then authentic, i.e. unencumbered by social pressure. (2) The lack of leakage about individual behavior protects the individual against potentially deleterious inferences.

dominates and transparency is in the aggregate preferable to safe spaces. In contrast, under high levels of divisiveness (one may have in mind high hostility, discrimination, verbal or physical violence), safe spaces allow one to be authentic without experiencing extreme hostility, and so a safe-space outcome socially dominates transparency. Section 3 also studies the behavioral impact of polarization (there are fewer moderates and more extremists; polarization is technically captured by a rotation of the distribution of types). Polarization raises the popularity of safe spaces, for both mechanical and incentive reasons. Section 3 then shows how to accommodate ancillary benefits, as when a member of a community of like-minded members finds better matches within this community. The section shows when equilibrium matching is endogamous (takes place within the safe space) and when it is exogamous. Finally, I allow agents to care more about the opinion of people who are akin to them, and show that such type-dependent reputational payoffs have some interesting implications.

Section 4 considers asymmetries in the distribution of preferences. Section 4.1 considers "one-sided polarization", under which only one of the sides becomes more radical. Right-wing individuals becoming more extremist interestingly raises the popularity of safe spaces on *both* sides. In contrast, a broader acceptance of right-wing ideas boosts the left-wing safe space, but contracts the right-wing one.

Section 4.2 studies the polar case of a binary action. While Sections 2 and 3 focus on the trinary action space with a symmetric preference distribution, the binary action version corresponds to very asymmetric preferences, with a majority preferring action $a = 0$ while the minority favors $a = 1$. The previous analysis applies with minor modifications. The same amalgam and reputational risk aversion operate.

Section 5 provides several extensions and applications, showcasing the richness of the framework. The first inquiry here concerns the robustness of the equilibrium to repeated interaction. I show that the safe space equilibrium (which again obtains for low hiding costs) is robust; in contrast, for high hiding costs, the resulting transparent equilibrium reveals more and more information over time, with partisanship unraveling in a Coasian manner. The second extension, noting that the very demand for privacy generates a cost of being outed by one's community, points out that our framework provides explanations for outings and coming outs, whose rationales would be elusive in a consensual-issue environment. Relatedly, the third extension analyses the case in which preserving privacy requires focusing on a social graph of like-minded peers whom one can trust.

Section 6 builds on the observation that, while safe spaces regroup like-minded individuals, their composition remains heterogeneous, creating scope for either additional signaling within the safe space or group split ups. Protected by the safety of the space, an agent may want to signal he is a true believer, which he would not do under transparency. Such signaling may range from a mere reduction in own utility (self-flogging) to tribalism (aggressions against out-group, terrorist acts, conspiracy theories, the spread of fake news, campus boycotts). Alternatively, the agent may be pressured by the community or the group's leadership to take an extremist stance. Section 6 shows both the potency of internal signaling and its limits (the possibility of a schism). Finally, Section 7 reviews the relevant literature and Section 8 concludes with avenues for future research. Missing proofs can be found in the Online Appendix.

2 Divisive behaviors

2.1 Baseline model

There is a mass 1 of agents, indexed by i, j. Each agent i picks an action $a_i \in \{-1, 0, +1\}$. Agent i can stay passive/neutral ($a_i = 0$) or act/pick a camp/be an activist ($a_i = -1$ or $+1$). Acting may involve a cost; let $c \geq 0$ denote this cost (time to participate in or demonstrating against an activity, cost of donating to a cause, etc.). Agent i has privately-known type (value, ideology) v_i and intrinsic motivation (non-image payoff from his action a_i):

$$v_i a_i - c|a_i|.$$

Preference heterogeneity. People disagree as to what is "moral" or "immoral", "good" or "bad", "right" or "wrong". The basic model posits a common knowledge cumulative distribution of tastes $F(v)$ on $\mathbb{R}$, unimodal and symmetric around 0 ($F(v) = 1 - F(-v)$ for all v). Its hazard rate is monotonic and the distribution has a mean (necessarily 0 given symmetry).

Privacy choice. Besides the choice of action a_i, an active agent i (for whom $|a_i| = 1$) faces a second choice: How much to disclose about his behavior. For expositional purposes, I invoke the passivity rationale to assume that moderates cannot prove their moderateness (I later note the equivalence with the duplicity rationale): There is nothing to disclose when not acting. An agnostic cannot prove he is not going to a religious office; a politically-neutral individual may not be able to demonstrate he is not carrying the card of the left- or right-wing party. I distinguish between the agent's *in-group*, composed of the agents adopting the same behavior (agent j such that $a_j = a_i$), and his *out-group* ($a_j \neq a_i$). I assume that:

- The agent's behavior is mechanically revealed to in-group agents, say as part of a shared activity within a facility. [Under the foundational assumptions made in Section 3, this assumption involves no loss of generality, as active agents want their peers to know and so share their information with them within a common space.]
- Active agent i hides his behavior from his out-group with probability $x_i \in \{0, 1\}$, the same for all agents in equilibrium; $x_i = 0$ means being transparent, while $x_i = 1$ is interpreted as joining a safe space of agents acting likewise.[6] Such hiding/acting

[6] More generally, what is needed for the results is that one's behavior be more visible to fellow adopters of the behavior (as is likely in most contexts).

covertly costs $h \geq 0$ to the agent.[7] [Again, there is no loss of generality given the assumptions made in Section 3, as the agent's preference in the absence of hiding cost will be not to disclose one's behavior to out-group members.[8]]

Agent i therefore can adopt one of three pure behaviors: passive agent ($a_i = 0$: behavior $\varnothing$), active, transparent agent ($|a_i| = 1$ and $x_i = 0$: behavior t), and active agent joining a safe space ($|a_i| = 1$ and $x_i = 1$: behavior s). Let $b_i \equiv \{a_i, x_i\}$ denote agent i's behavior. Given the symmetry of the model and the monotonicity of the payoff in type, I look for a symmetric equilibrium $\{v^*, x\}$, where (a) agents with $v_i \geq v^* \geq 0$ select $a_i = +1$, those with $v_i \leq -v^*$ select $a_i = -1$, and the others, if any, remain passive; and (b) a fraction x of active agents hide their behavior from their out-group. [Later I provide sufficient conditions for there to be a unique equilibrium.]

Image payoffs. Agent i's action and disclosure strategy determines other agents' beliefs about his type. At this stage, I posit a general, audience-contingent, reputational/image payoff. Namely, let $R_{b_i}^{v_i}(v^*, x)$ denote agent i's payoff when having type v_i and choosing behavior b_i, when equilibrium behavior is summarized by $\{v^*, x\}$. The utility of agent i with type v_i and behavior $b_i = \{a_i, x_i\}$ is therefore

$$U_i(v_i, b_i | v^*, x) = v_i a_i - [c + hx_i]|a_i| + R_{b_i}^{v_i}(v^*, x).$$

An equilibrium is as usual a fixed point: Given the population's behavior $\{v^*, x\}$, type-v_i agent's maximization with respect to behavior b_i, for all v_i, gives rise to aggregate behavior $\{v^*, x\}$ in the population. Let us assume that the cross-partial derivative of $\max_{b_i} U_i(v_i, b_i | v^*, x)$ with respect to v_i and a_i is positive, guaranteeing monotonicity of a_i in v_i (this will be the case in particular if $R_{b_i}^{v_i}$ is type-independent, or if the intensity of image concerns is not too large). For conciseness, let $R_{b_i}(v^*, x) \equiv R_{b_i}^{v^*}(v^*, x)$ denote the cutoff agent's image payoff (also equal to the payoff of agent $-v^*$ by symmetry). Note that image payoffs under transparent behavior, including the cutoff's $R_t(v^*, x)$, do not depend on x: All other agents then know that the agent has picked $a_i = 1$ and so that his type satisfies $v_i \geq v^*$; I will therefore abuse notation and label $R_t(v^*)$ this image payoff. Let

$$\Delta_t(v^*, x) \equiv R_t(v^*) - R_\varnothing(v^*, x)$$

denote the cutoff type's *image incentive to act transparently rather than not acting*, and

$$\Delta_s(v^*, x) \equiv R_s(v^*, x) - R_\varnothing(v^*, x)$$

[7] Technological progress alters the cost h of hiding in opposite ways. On the one hand, some of our formerly private spaces (e.g., being in the anonymity of a large crowd) are transformed into public spaces by smartphones, AI or smart glasses. On the other hand, technology may restore incentives to engage by enabling the creation of new spaces of like-minded individuals, most notably within social networks (with potentially dire consequences, though, as we discuss in Section 6).

[8] Our requirement that disclosure is uniform across the out-group involves no loss of generality. First, note that the agent is unable to distinguish among out-group agents if all active agents conceal their behavior. Second, if this is not the case, the agent de facto faces two out-groups: Those whom he knows to have acted transparently, and those about whom he has no information. Nonetheless, under the assumptions of Section 3, the desired disclosure choice -i.e. no disclosure in the absence of hiding cost- is the same for both out-groups even if the agent can selectively disclose.

denote the cutoff type's *image incentive to act in a safe space (covertly) rather than not acting.* Using these two definitions I can define the cutoff type's *incentive to hide when acting* as

$$\Delta_s(v^*, x) - \Delta_t(v^*, x) = R_s(v^*, x) - R_t(v^*).$$

I make the following assumptions:

Assumption 1 *(image incentive to act). For all (v^*, x),*

(i) acting transparently (resp. covertly) reduces (increases) the agent's image payoff relative to not acting: $\Delta_t(v^*, x) \leq 0 \leq \Delta_s(v^*, x)$

(ii) incentives to act transparently or covertly, $\Delta_t(v^*, x)$ or $\Delta_s(v^*, x)$, increase with the fraction x of active agents who hide.

Assumption 2 *(image incentive to hide). For all (v^*, x),*

(i) hiding strictly increases the agent's image payoff ($\Delta_s(v^*, x) - \Delta_t(v^*, x) > 0$)

(ii) the incentive to hide, $\Delta_s(v^*, x) - \Delta_t(v^*, x)$, decreases with activism (increases with v^*) and decreases with the fraction x of active agents who hide.

These assumptions capture the following ideas:

- *Reputational risk aversion: The cost of spooking the out-group dominates the benefit of reassuring the in-group.* This effect speaks to parts (i) of the two assumptions. Relative to not acting, acting covertly does not alter the information (none) that the out-group receives about the agent. So acting covertly only serves to reassure the in-group ("*I am one of yours*"), and therefore $\Delta_s(v^*, x) \geq 0$. Relative to not acting, acting transparently reveals dissonance with the out-group and congruence with the in-group. The assumption that $\Delta_t(v^*, x) \leq 0$ means a transparently-active agent makes on average a worse impression than a passive one (the qualifier "on average" is important here as he will have a better image within his in-group). That $\Delta_s(v^*, x) - \Delta_t(v^*, x) > 0$ reflects the idea that the active agent strictly prefers the out-group not to know (the in-group is always informed, whether the behavior is transparent or covert).

- *Amalgam effect.* The amalgam effect, which speaks to parts (ii) of the two assumptions, refers to a negative externality exerted by agents acting covertly onto passive ones. When active agents conceal their behavior from their out-group, these out-group agents suspect that agents they do not know anything about may well be extremists mascarading as moderates. This reduces the reputation payoff from not acting, and thus increases the incentive to act, whether covertly or transparently (Assumption 1(ii)). Finally, the assumption that $\Delta_s(v^*, x) - \Delta_t(v^*, x)$ decreases with x reflects the idea that an active agent's strategy of hiding from his out-group is more effective when other active agents are transparent; the hiding agent then builds a solid reputation for being a moderate. The amalgam effect makes hiding behaviors strategic substitutes.

To guarantee uniqueness of the cutoffs, I make the technical assumption that the cutoff's incentive to act overtly for a given x is increasing in v^*:

Assumption 3 *(monotonicity).*

(i) The cutoff's equilibrium incentive to act transparently, $T(v^*, x) \equiv v^* - c + \Delta_t(v^*, x)$, *is strictly increasing in* v^*.

(ii) $\Delta_t(0, 0) = lim_{v^* \to 0} \Delta_t(v^*, 0)$.

Assumptions 2(ii) and 3 combined further imply that the cutoff's equilibrium incentive to act covertly, $S(v^*, x) \equiv v^* - c + \Delta_s(v^*, x)$, is also strictly increasing in v^*. A sufficient condition for Assumption 3 to hold is that image concerns (the parameter μ of intensity of image concerns in Section 3) not be too large.

A comment is in order regarding part (ii) of Assumption 3. $\Delta_t(0, 0)$ is not well-defined, as the absence of observed activism in a transparent equilibrium has probability 0 if $v^* = 0$. Part (ii) thus resolves the off-path indeterminacy of beliefs. This selection will guarantee a unique equilibrium. Under more pessimistic beliefs, the configuration $\{v^* = 0, x = 0\}$ may coexist with that derived below. As this possibility offers no new insight, I ignore it for conciseness by making Assumption 3 (ii).

Equilibria. The equilibrium conditions are:

Safe space equilibrium $(x = 1)$

$$v^* - c + \Delta_s(v^*, 1) = h \tag{1}$$

$$\Delta_s(v^*, 1) - \Delta_t(v^*, 1) \geq h$$

Mixed-strategy equilibrium $(0 < x < 1)$

$$v^* - c + \Delta_s(v^*, x) = h \tag{2}$$

$$\Delta_s(v^*, x) - \Delta_t(v^*, x) = h$$

Transparent equilibrium $(x = 0)$

$$v^* - c + \Delta_t(v^*, 0) = 0 \tag{3}$$

$$\Delta_s(v^*, 0) - \Delta_t(v^*, 0) \leq h$$

Comparison with full privacy and full transparency. Let us compare the agents' actions with the ones that would prevail if no information about their behavior filtered out, even to the in-group (full privacy). In that case, the agents would behave "autonomously", that is according to their own preferences, without any social pressure: $v^* = c$. Assumption 1(i) implies that the cutoff in a transparent equilibrium, which I will label v^t, satisfies $v^t \geq c$: Agents are fearful of acting if their behavior is disclosed to everyone. The same assumption also implies that in the absence of hiding cost, the equilibrium, which is necessarily a safe-space one, involves much activism as active agents can costlessly reassure their in-group without incurring any reputational penalty from the out-group: The cutoff v^s satisfies $v^s \leq c$. This cutoff may be equal to 0: A strong amalgam effect forces agents to take sides. v^s is strictly positive if and only if $c > R_s(0,1) - R_{\varnothing}(0,1)$. This condition guarantees that there are passive agents even for $h = 0$. If it fails, then $v^s = 0$ over an interval $h \in [0, h_0]$.

The comparison between involvement under transparent and safe-space behavior also turns the standard result for consensual behaviors that a higher visibility increases compliance on its head: When behaviors are divisive, privacy encourages activism. For consensual behaviors transparency makes high types invest in reputation to separate themselves from low types. Divisiveness kills this incentive as what is approved by some is (more strongly) frowned upon by others; by contrast, signaling a high type (in absolute value now) remains valuable if the information is shared only among like-minded peers, which requires the privacy of a safe space.

Proposition 1 *(equilibrium). Under Assumptions 1-3, there exists a unique symmetric equilibrium and this equilibrium is characterized by two cutoffs* $0 < h_1 < h_2$. *As the hiding cost* h *increases, the equilibrium involves first safe spaces (for* $h \in [0, h_1]$*), then a mixed strategy (for* $h \in (h_1, h_2)$*), and finally transparency (for* $h \geq h_2$*). Agents are less inclined to act as hiding becomes more difficult:* v^**increases with* h.

Proof of Proposition 1. T is increasing in its first argument (Assumption 3) as well as its second (Assumption 1(ii)). Let $\mathcal{V}(x)$, a decreasing function, be defined by $T(\mathcal{V}(x), x) = 0$. A transparent equilibrium exists if and only if $T(\mathcal{V}(0), 0) \equiv h_2 \leq h$. Next, $\Delta_s(\mathcal{V}(x), x) - \Delta_t(\mathcal{V}(x), x)$ is decreasing in x (Assumption 2(ii)). Thus a mixed-strategy equilibrium exists if and only if $\Delta_s(\mathcal{V}(1), 1) - \Delta_t(\mathcal{V}(1), 1) \equiv h_1 < h < h_2 \equiv \Delta_s(\mathcal{V}(0), 0) - \Delta_t(\mathcal{V}(0), 0)$. Finally, a safe-space equilibrium ($x = 1$) exists if and only if $h \leq h_1$. ■

Online Appendix A shows that, for reputational payoffs arising from bilateral reputations (Section 3), there is no asymmetric equilibrium. The equilibrium is therefore unique.

The duplicity rationale. Assume in contrast that (i) all actions $a \in \{-1, 0, +1\}$ are observed unless (costlessly) concealed by the agent, and (ii) an agent can choose either two actions[9] at (psychological or time) cost h or a single action at no cost. Because actions can be disclosed, the moderates escape the amalgam effect only if all agents select a single action. The condition for a transparent equilibrium -here the equilibrium without duplicity- is again given by (3). In contrast, a right-leaning agent, say, can choose both

[9] Excluding an engagement in the three actions involves no loss of generality.

action $a = 0$ (and disclose it publicly) and action $a = 1$ (and disclose it only to the in-group). The equilibrium conditions for a safe and mixed-strategy equilibrium are given by equations (1) and (2), respectively.

3 Bilateral reputations foundations

This section builds an agent's overall image payoff from bilateral reputations and posits that agent i with type v_i values his reputational payoff $r(\hat{v}_{ji}, v_j|v_i)$ with agent j, where $\hat{v}_{ji}$ denotes the expected type of agent i conditional on whatever agent j observes about his behavior.[10] Thus, reputation is in the eye of the beholder, in two related ways: First, unlike in standard models of prosocial behavior or of conformity, agents in the audience differ in their appreciations of a fellow agent's behavior: Agent i's reputation depends not only on his behavior, but also on agent j's type v_j. Second, the visibility of agent i's action to agent j depends endogenously on the latter's type, even though this type is private information: r depends on $\hat{v}_{ji}$. Agent i manipulates the visibility of his behavior in an audience-contingent manner. Finally, the reputational payoff may or may not depend on the agent's type v_i. The overall reputational payoff of agent i adds up bilateral payoffs:[11]

$$R_i = R_{b_i}^{v_i}(v^*, x) \equiv \int_{-\infty}^{+\infty} r(\hat{v}_{ji}, v_j|v_i) dF(v_j).$$

Until Section 3.5, I will focus for expositional simplicity on type-independent reputational payoffs, in which $r(\hat{v}_{ji}, v_j \mid v_i)$ does not depend on v_i. This is a reasonable assumption when individual i only cares about avoiding bad treatments by others (negative comments, violence, discrimination) and eliciting friendly ones (in contrast, r depends on v_i if for example agent i cares more about the opinion of the in-group than of the out-group: See Section 3.5). By a slight abuse of notation, I will then denote the payoff $r(\hat{v}_{ji}, v_j)$. I assume that r is twice continuously differentiable and denote $r_1 \equiv \partial r/\partial \hat{v}$, $r_{11} \equiv \partial^2 r/\partial \hat{v}^2$, etc.

[10] We thus anchor agent i's reputational payoff vis-à-vis agent j to his *expected* type given j's information about his behavior. In that sense, agent i is viewed as representative of a perceived group. This paradigm seems relevant in a number of contexts, as an agent may be concerned about how other agents will treat him, and this treatment will depend on the average opinion that the other agents have of him. An alternative hypothesis, entertained in the Online Appendix G, is that agent i's reputational payoff corresponds to a perception by agent j of agent i as a *random* member of the perceived group. The two formulations coincide when the function r is linear in its first argument (the special case of a "positional image" below), but differ in general. The equilibrium and welfare characterizations for this second hypothesis however are identical to those of the positional image case even if the function r is not linear in its first argument.

[11] With perceived type $\hat{v}_{ji}(b_i, v_j|v^*, x)$, this notation is shortcut for:

$$R_i = R_{b_i}^{v_i}(v^*, x) = \int_{-\infty}^{+\infty} r\big(\hat{v}_{ji}(b_i, v_j), v_j)|v^*, x), v_j|v_i\big) dF(v_j).$$

The reputation function r is assumed to satisfy three reasonable properties besides symmetry: First, the agent ceteris paribus wants to limit perceived taste dissonance with his audience; second, perceived ideological differences come at an increasing marginal cost (r is weakly concave in its first argument); and third, an agent gains from being perceived by an activist group's members as representative of the group rather than as the average type in the entire population. These properties are satisfied by a range of models, including two simple ones: The first involves a positional image model, in which reputation acquisition is a constant-sum game and v_j affects the (positive or negative) weight on reputation $\hat{v}_{ji}$. The second takes r to depend negatively on the distance between $\hat{v}_{ji}$ and v_j, as measured by the L^p-norm. Finally, I specialize Assumption 3 to this model, guaranteeing the uniqueness of the (symmetric) cutoffs.

Assumption 4 *(image concerns under bilateral image concerns). The bilateral reputational payoff $r(\hat{v}, v)$, where $\hat{v}$ is the agent's perceived type in the eye of the beholder v, is assumed to satisfy:*

(i) Symmetry: For all $(\hat{v}, v)$, $r(-\hat{v}, -v) = r(\hat{v}, v)$.

(ii) Distaste for dissonance: Ceteris paribus, agents want to ingratiate themselves with others. Suppose that $v > 0$.[12] Then for all $\hat{v} < v$,[13] $r_1(\hat{v}, v) > 0$.

(iii) Concavity (reputational risk aversion): Perceived ideological disapproval has an increasing marginal cost: For all $(\hat{v}, v)$, $r_{11}(\hat{v}, v) \leq 0$.

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

(v) Monotonicity (unique cutoff). Letting $M^+(v^*) \equiv E(v \mid v \geq v^*)$ and $M^-(v^*) \equiv E[v \mid v \leq v^*]$ denote the left- and right- truncated means at cutoff v^*, the incentive to act transparently

$$v^* - c + \int_{-\infty}^{+\infty} r(M^+(v^*), v)dF(v) - \int_{-v^*}^{v^*} r(0, v)dF(v) - 2\int_{v^*}^{+\infty} r(M_x^-(v^*), v)dF(v)$$

is strictly increasing in v^* for all x, where $M_x^-(v^*)$ is the right-truncated mean when

[12] By symmetry, Assumption 4(ii) implies that for $v < 0$ and $\hat{v} > v$, then $r_1(\hat{v}, v) < 0$. To see this, use Assumption 4(i). Because $r(\hat{v}, v) = r(-\hat{v}, -v)$, then $r_1(\hat{v}, v) = -r_1(-\hat{v}, -v)$. When $v < 0$, $-v > 0$, and if $\hat{v} > v$, $-\hat{v} < -v$. Assumption 4(ii) then implies that $r_1(-\hat{v}, -v) > 0 \Leftrightarrow r_1(\hat{v}, v) < 0$.

[13] Note that I make no assumption regarding r_1 for $\hat{v} > v > 0$ (or symmetrically for $\hat{v} < v < 0$). Indeed, in illustration 1 (resp. illustration 2) below, $r_1(\hat{v}, v) > 0$ (resp. < 0) for $\hat{v} > v > 0$.

hiding has probability x:[14]

$$M_x^-(v^*) \equiv \frac{xF(-v^*)M^-(-v^*)}{xF(-v^*) + [F(v^*) - F(-v^*)]}.$$

3.1 Illustrations

In all illustrations below, the image concerns will be scaled by an image intensity parameter $\mu \geq 0$. They satisfy Assumption 4 provided μ is not too large (so Assumption 4(v) is satisfied). These illustrations are selected not only for their tractability, but also for their different properties.

Illustration 1: Positional image. Suppose that

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

where the audience-contingent weight $\theta(v)$ is an increasing and antisymmetric function ($\theta(-v) = -\theta(v)$), necessarily satisfying $\theta(0) = 0$. Note that the concavity property (Assumption 4(iii)) is satisfied with equality ($r_{11} = 0$). Indeed, $r_{11} = 0$ is characteristic of the positional image paradigm.[15] Such image concerns are called positional or zero-sum, because total reputation in society is fixed:

$$E_{\{v_i,v_j\}}[\mu\theta(v_j)\hat{v}_{ji}] = 0.$$

The positional-image model assumes that $a_i = +1$ is frowned upon by those who view this behavior as reprehensible ($v_j < 0$), and the more so, the more opinionated j is and the more extremist agent i is perceived to be (the higher $\hat{v}_{ji}$ is); by contrast, if observed, this action boosts agent i's reputation with those who approve of this action ($v_j > 0$), again the more so the more approbative the audience and the higher the perceived faith of agent i. And conversely for action $a_i = -1$.

Illustration 2: Placating image concerns. Suppose that agents want to be perceived as close in values as possible to their audience.[16] Let such concerns be measured by the F-norm corresponding to the L^p-norm distance between the agent's perceived type and the beholder's type:

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

It is homogenous of degree $p \geq 1$ (for example, the Euclidean distance corresponds to $p = 2$: $r(\hat{v}, v) = -\mu(\hat{v} - v)^2$). The requirement $p > 1$ reflects the idea that the agent cares more about hostile opinions than about favorable ones, say because hostile opinions may

[14] Note that $M_0^-(0)$ is not well-defined, as the observation of $a = 0$ in a transparent equilibrium has probability 0 if $v^* = 0$ (furthermore $lim_{v^* \to 0} M_0^-(v^*) = 0$, while $lim_{x \to 0} M_x^-(0) = M^-(0)$). Assumption 3 (ii) implies that $M_0^-(0) = 0$. See the discussion following Assumption 3.

[15] Suppose $r_{11} = 0$. Then there exist $\theta(v)$ and $\gamma(v)$ such that $r(\hat{v}, v) = \theta(v)\hat{v} + \gamma(v)$. Assumption 4(i) implies that γ is symmetric and θ antisymmetric. Image is therefore constant-sum.

[16] This is not the case for a positional image: If $v_j > 0$, agent i wants $\hat{v}_{ji}$ to be as high as possible.

trigger hate and violence. These image concerns trivially satisfy Assumption 4(i), (ii) and (iii) for all p. More interestingly, they also satisfy part (iv) of Assumption 4 for all p, as is shown in Online Appendix D.

Online Appendix E analyses non-additive forms:

Illustration 3. The "true L^p norm" corresponds to overall reputational payoff

$$R_i \equiv -\mu \left(\int_{-\infty}^{+\infty} |\hat{v}_{ji} - v_j|^p dF(v_j) \right)^{\frac{1}{p}}.$$

Illustration 4: Maximum norm. Suppose that the support of F is finite on $[-V, +V]$. The maximum norm is obtained by taking the limit as $p \to +\infty$ of the true L^p norm, so

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

That is, the agent's reputational concerns focus on the most hostile (in the sense of perceived distance) member of his audience. The maximum norm captures an extreme form of conflict aversion.

Proposition 2 *(bilateral image concerns). An agent's overall image payoff $R^{v_i}_{b_i}(v^*, x)$ satisfies Assumptions 1-3 if it aggregates bilateral image payoffs satisfying Assumption 4. This is indeed the case for the four illustrations above, provided that the intensity of image concerns μ is not too large (so that Assumption 4(v) is satisfied).*

3.2 Polarization and safe spaces

I next study the impact of a change in the distribution F of types.

Definition 1 *(polarization). Let $F(v;\rho)$ denote a family of unimodal, symmetric distributions, smooth jointly in (v,ρ). The population becomes "more polarized" if $\rho \in \mathbb{R}$ indexes a rotation with 0 as rotation point: $F_\rho < 0$ for $v > 0$ (so by symmetry $F_\rho > 0$ for $v < 0$).*

To illustrate the impact of polarization concisely, I focus on a safe-space equilibrium (h low enough) and either a positional image or the F-norm corresponding to the L^p norm.

Proposition 3 *(polarization). Suppose a safe-space equilibrium (h low enough), either a positional image or the F-norm corresponding to the L^p norm, and that the hazard rate of F is monotone. A symmetric increase in polarization (in ρ) modifies behavior (v^*increases) and increases the fraction of active agents in a safe space, $2[1-F(v^*(\rho);\rho)]$, both mechanically and through behavior change.*

Proof of Proposition 3. The equilibrium cutoff for a safe-space equilibrium, v^*, is given by

$$v^* - c + \int_{v^*}^{+\infty} \big[r(M^+(v^*;\rho), v) - r(M^-(v^*;\rho), v)\big] dF(v;\rho) = h. \tag{4}$$

Note that for both the positional image and the F-norm corresponding to the L^p norm, a higher image is more valued, the higher the audience's type: $r_{12} > 0$. Given the monotonicity Assumption 4(v), differentiating (4), integrating by parts, and using $F_\rho(v^*) < 0$ for $v^* > 0$ (rotation) and $F_\rho(+\infty) = 0$ yields

$$\text{sgn}\left(\frac{dv^*}{d\rho}\right) = \text{sgn}\left(\Gamma_1(v^*) + \Gamma_2(v^*)\right),$$

where $\Gamma_1(v^*) \equiv \int_{v^*}^{+\infty}[r_2(M^+(v^*), v) - r_2(M^-(v^*), v)]F_\rho(v^*) < 0$ (since $r_{12} > 0$). Let us show that $\Gamma_2(v^*) \equiv [r(M^+(v^*), v^*) - r(M^-(v^*), v^*)]F_\rho(v^*) < 0$, knowing $F_\rho(v^*) < 0$.

Consider first the L^p norm: $r(M^+(v^*), v^*) - r(M^-(v^*), v^*) = \mu[(v^* - M^-(v^*))^p - (M^+(v^*) - v^*)^p]$. Under a monotone hazard rate, $(M^+(v^*))' \in (0, 1)$ and $(M^-(v^*))' \in (0, 1)$ (An 1998). Assumption 4(i) (symmetry) imposes that $\Gamma_2(0) = 0$. So $\Gamma_2(v^*) > 0$ for $v^* > 0$. This implies that $dv^*/d\rho < 0$.

Consider next a positional image. The increase in the weight put in the tails ($F_\rho < 0$ for $v^* > 0$) again mechanically raises the number of activists for a given v^*. This is the only effect under transparency as $v^* = c$. Let $\Delta(v^*) \equiv M^+(v^*) - M^-(v^*)$ and $\Theta(v^*; \rho) \equiv \mu \int_{v^*}^{+\infty} \theta(v) dF(v; \rho)$. Under a safe-space equilibrium, $v^* = v^s(\rho)$ is given by

$$v^s(\rho) - c + \Theta(v^s(\rho); \rho)\Delta(v^s(\rho); \rho) = h. \tag{5}$$

An increase in polarization changes both $\Theta(v^s; \rho)$ and $\Delta(v^s; \rho)$. It increases $\Delta(v^s; \rho)$ from Adriani-Sonderegger (2019)'s Proposition 3 (which states that a mean-preserving spread increases Δ; because the rotation point is the mean - 0 -, the rotation is a mean-preserving spread). As for $\Theta(v^s; \rho)$, an integration by parts yields $\frac{\partial}{\partial\rho} \int_{v^s}^{+\infty} \theta(v) dF(v; \rho) = -\int_{v^s}^{+\infty} \theta'(v) F_\rho(v; \rho) dv - \theta(v^s) F_\rho(v^s; \rho) \geq 0$. So both effects go in the same direction, and v^s decreases with ρ. Participation ($|a_i| = 1$) also increases as $\frac{\partial}{\partial\rho}[1 - F(v^s(\rho); \rho)] = -F_\rho - f\frac{dv^s}{d\rho} > 0$. Intuitively, as right-wingers become more opinionated, the opinion of the right-wing in-group matters more ($\Theta(v^s; \rho)$ increases). Furthermore, the perceived type differential ($\Delta(v^s; \rho)$) between right-wing activists and their outgroup increases (fewer moderates and more extremists). ■

3.3 Welfare

This section performs a welfare analysis by aggregating utilities in the population. Let's issue two caveats before doing so. First, there are specific reasons to be wary of the utilitarian approach in this context. For instance, the protection of minorities is viewed from the specific angle of their incurring the disdain or the wrath of others who think differently. Concepts like inclusion and dignity may, but need not fit with this perspective. Second, there may be incidental costs and benefits that are either caused by, or orthogonal to the emergence of safe spaces studied here. For example, viral contagion effects should be added to our welfare analysis if the divisive issue is the attitude toward vaccines.[17] An example of social benefit of homophilic regrouping is that people can meet other individuals whose company they enjoy (gays may find a suitable partner in a gay bar, political militants can enjoy discussions with other militants), independently of safety concerns. Such ancillary benefits can be incorporated into our analysis as the next subsection demonstrates.

Full privacy benchmark. An interesting point of comparison is the polar case in which no one observes the agent's behavior, and so there are no social image concerns.[18] Let us define the "authentic self" as the behavior that would prevail under such full privacy ("fp"): The cutoff would then be $v^* = v^{fp} = c$.

Social pressure externalities. Consider the welfare of passive agents in a safe-space equilibrium with cutoff $v^* = v^s$:

$$R_\varnothing(v^*, 1) \equiv 2\int_{v^s}^{+\infty} r(M^-(v^s), v)dF(v) + \int_{-v^s}^{v^s} r(0, v)dF(v).$$

This payoff is lower than the one, equal to $\int_{-\infty}^{+\infty} r(0,v)dF(v)$, they would obtain either under transparency or under full privacy. More generally, the amalgam effect (which more broadly exists if and only if $x > 0$) reduces the passive agents' welfare relative to the case of full privacy or of transparency.

Aggregate image and welfare. We start by noting that the total image payoff over the entire population and welfare are maximized under full privacy. Let $\mathcal{R} \equiv \int_{-\infty}^{+\infty} [\int_{-\infty}^{+\infty} r(\hat{v}_{ji}, v_j)dF(v_j)]$ $dF(v_i)$ be the total reputational payoff over the entire population. We denote by $\mathcal{R}^{fp} \equiv \int_{-\infty}^{+\infty} r(0, v)dF(v)$, $\mathcal{R}^s(v^*)$, $\mathcal{R}^m(v^*, x)$, and $\mathcal{R}^t(v^*)$ its realizations under full privacy, safe spaces, mixed region, and transparency ($\mathcal{R}^s(v^*) = \mathcal{R}^m(v^*, 1)$ and $\mathcal{R}^t(v^*) = \mathcal{R}^m(v^*, 0)$). We also consider the thought experiment of "full transparency" (ft), in which the agent's *type* is revealed to all ($\mathcal{R}^{ft} \equiv \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} r(\tilde{v}, v)dF(v)\, dF(\tilde{v})$).

Because, in all these configurations, the cutoff v^* is by definition indifferent between $a_i = 0$ and $a_i = 1$, welfare in any equilibrium configuration can be written as the cutoff type's payoff when remaining passive, which is their reputational payoff $R_\varnothing(v^*, x)$, plus the rent of more committed types:

$$W(v^*, x) = R_\varnothing(v^*, x) + 2\int_{v^*}^{+\infty} (v - v^*)dF(v).$$

The proof of the following proposition can be found in Online Appendix B.[19]

[17] We could also add potential benefits for the audience of having more information; such benefits would tilt the balance in favour of transparency.

[18] While I can think about activities (such as being deep in our thoughts) in which full privacy can be enjoyed, for most activities it is not clear that the individual can or wants to engage in it in a non-social manner. Even sexuality, practiced in the secrecy of the home, has strong social components (finding partners, enjoying the public space/a normal life with the partner). Similarly, while we can keep our political and social views for ourselves, we enjoy sharing them with others. Evolution has made humans a deeply social species.

[19] The ranking among full privacy, full transparency, and the three equilibrium configurations in Propo-

Proposition 4 *(total image payoff and welfare). Under type-independent image payoffs satisfying Assumption 4,*

- *(i) More information reduces total image payoff: For all v^* (and $x \in (0,1)$)*

$$\mathcal{R}^{fp} \geq \mathcal{R}^{s}(v^*) \geq \mathcal{R}^{m}(v^*, x) \geq \mathcal{R}^{t}(v^*) \geq \mathcal{R}^{ft}$$

with strict inequalities when $r_{11} < 0$, and equalities when $r_{11} \equiv 0$.

- *(ii) Full privacy, which furthermore generates authentic behavior, yields an upper bound on equilibrium welfare:*

$$W^{fp} \geq \max\{W^s, W^m, W^t\},$$

strictly so unless $r_{11} = 0$ (in which case $W^t = W^{fp}$), where W is the welfare when parameters lead to regime $r \in \{s, m, t\}$.

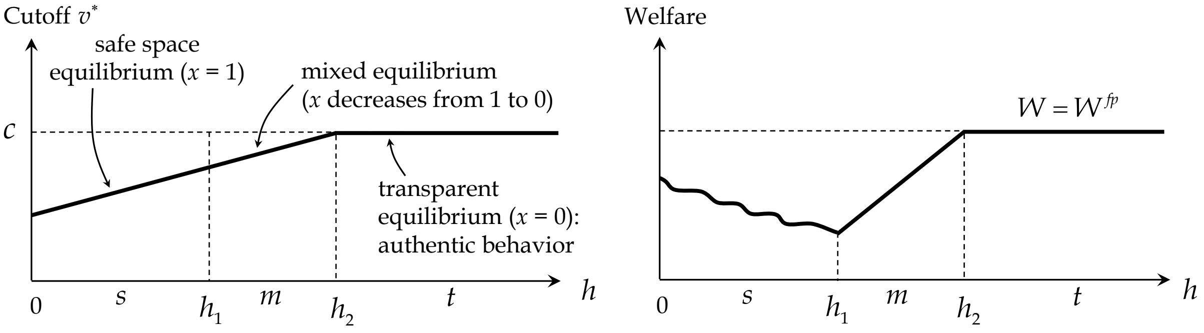

Figure 1: Cutoffs and welfare under a positional image

Finally, Online Appendix C looks at the cases of positional and maximum norm image concerns, which are somehow polar cases and are depicted in Figures 1 and 2. The positional-image and maximum-norm illustrations can be summarized in the following way:

Positional image concerns. Suppose that $r(\hat{v}, v) = \mu\theta(v)\hat{v}$.

- Authenticity. Safe spaces, by encouraging agents to impress like-minded peers, do not promote authentic behavior. The authenticity in the safe-space equilibrium (which exists if and only if $h \leq h_1$) decreases with the importance of social approval.

sition 4 extends to asymmetric distribution functions F. In contrast, the concavity assumption $r_{11} \leq 0$ is central to the normative property that the full-privacy allocation would be optimal if it were incentive compatible, and more broadly to welfare comparisons. The positive analysis need not be altered under a convex reputation function ($r_{11} > 0$), though. To see why, consider for instance the positional-reputation case ($r_{11} = 0$). For h low, the agents in a safe space strictly prefer hiding from the out-group. This is still true as long as the reputation function is not too convex. Put differently, the amalgam/reputation-stealing effect on moderates still prevails.

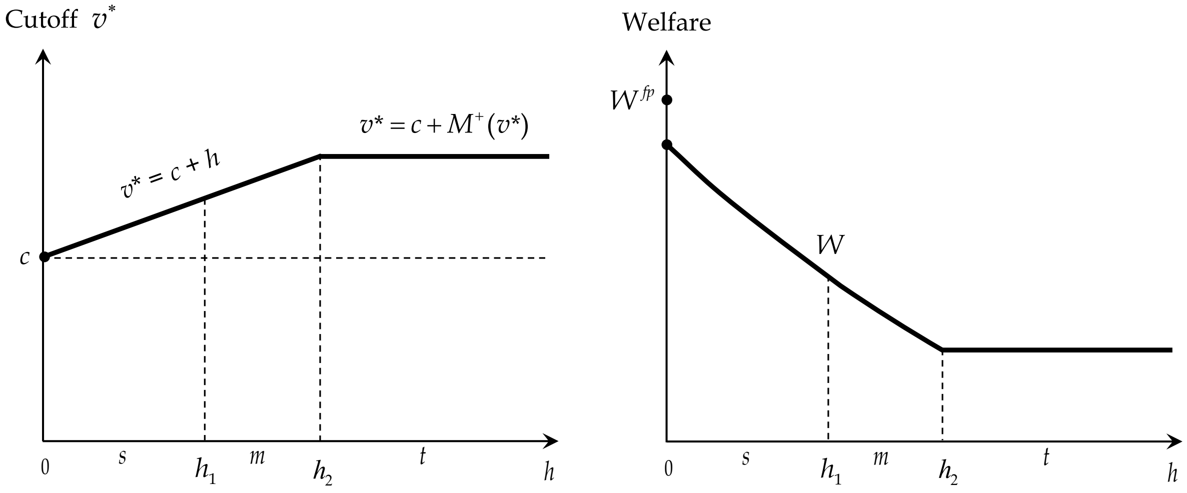

Figure 2: Cutoffs and welfare under the maximum norm

- Welfare. Welfare is highest under transparency ($h \geq h_2$): Image is a zero-sum game and transparency both eliminates the negative externality on passive agents and promotes authenticity.

The maximum norm. Suppose that $R_i = -\mu \max_j |\hat{v}_{ji} - v_j|$,

- Authenticity. The level of activity is always below the authentic level ($v^* > c$ if $\mu > 0$).
- Welfare. There exists $\mu_0 > 0$ such that for all $\mu \leq \mu_0$, the cutoff is continuously increasing in h, from $v^* = c$ to $v^* = c + M^+(v^*)$. Welfare is continuously decreasing in h.

3.4 Ancillary benefits of safe spaces

Safety is not the only benefit offered by communities made of similar agents. Groups of peers may presumably solve coordination problems; it is easier for a gay person to find a suitable partner in a gay bar. Furthermore, agents with similar interests on a divisive social issue (abortion rights, for example) may share other values and enjoy being with each other. Would such ancillary benefits lead to different welfare implications?

Online Appendix F studies how ancillary benefits from belonging to a safe space can be incorporated into my analysis. I assume that the ancillary benefit comes from matching and that the matching benefit decreases with ideological distance. An agent, beyond choosing a, may match (randomly) within one of three groups: Within a safe space that emerges in equilibrium (for this, the agent must have joined the safe space), or within the "general pool" that potentially admits those who have revealed nothing to members of this general pool (by being either active within a safe space or passive). Passive agents necessarily match within the general pool, while active agents who have concealed their behavior to the out-group choose to match either in the safe space ("*endogamous matching*") or in the general pool ("*exogamous matching*").

Starting with the obvious, adding this ancillary option benefit of homophilic groups ceteris paribus weakly lowers the value of the cutoff $v^* = v^s$ where v^s is determined by (4) (or (1)), as the larger option set for matching increases the benefit of being active. More interestingly, exogamous matching arises for the most moderate activists when safe spaces are attractive (say, because the costs c and h are small); the cutoff v^* is then more at ease matching with a moderate type than with an activist one. The cutoff type then does not enjoy a dual benefit of being in a safe space, and the equilibrium condition is still given by condition (4). In contrast, when endogamous matching is the pattern for all members of the safe space, the cutoff v^* is smaller than the value given by (4). Overall, the previous analysis accommodates ancillary benefits, without altering its qualitative insights. In particular:

Suppose that (i) the matching benefit decreases linearly with the expected ideological distance with the (random) match; (ii) the hazard rate of distribution F is monotonic, (iii) h is sufficiently small so that, in the absence of matching benefit, the equilibrium is a safe-space one. Then matching is exogamous if $c < c^\#$ and endogamous for $c > c^\#$, where $c^\# > 0$ is such that $M^+(v^s) = 2v^s$ (recall that v^s solves (4)).

3.5 Type-dependent reputational payoffs

One can also think of cases in which reputational payoffs are type-dependent. Agent i may care more about the opinion of people who are akin to him. The analysis accommodates the more general case of type-dependent reputational payoffs as long as μ is not too large.[20]

There is one interesting twist, though: The disclosure of one's behavior may not be uniform within an in-group of agents adopting the same behavior. To see this, suppose that agents care more about the perceptions of agents who are similar to them:

$$r(\hat{v}_{ji}, v_j | v_i) = \mu \Phi(\hat{v}_{ji}, v_j) \Lambda[|v_i - v_j|]$$

where Φ satisfies Assumption 4 and $\Lambda' < 0$. And consider a safe-space equilibrium with cutoffs $\{-v^*, v^*\}$. The activists all hide if and only if for all $v_i \geq v^*$

$$\begin{aligned} h \leq \mu \Big[\int_{-\infty}^{-v^*} \big[\Phi(M^+(-v^*), v_j)) - \Phi(M^+(v^*), v_j)\big] \Lambda(v_i - v_j) dF(v_j) \\ + \int_{-v^*}^{v^*} \big[\Phi(0, v_j) - \Phi(M^+(v^*), v_j)\big] \Lambda(v_i - v_j) dF(v_j) \Big]. \end{aligned}$$

[20] The existence of an ordered behavior is not guaranteed otherwise: Consider, e.g., the -unlikely- case in which a left-winger wants to appeal to right-wingers so much that he picks the right-wing action.

In the set of types above v^*, a higher type v_i cares less about the perception of the out-group (the RHS of this inequality is decreasing in v_i). And so they are more willing to go transparent to avoid incurring hiding cost h. Of course, the equilibria are still safe-space equilibria for low hiding costs; but new equilibrium configurations can arise for intermediate hiding costs: *Agents with extreme convictions do not mind being transparent as they care little about their image with out-group agents. Thus, activists separate between moderates, who hide from the out-group, and extremists, who do not.*

4 Asymmetric distributions and binary actions

This section shows that the model can fruitfully be applied to asymmetric distributions. Under monotonicity of the payoff functions, incentive compatibility implies that an equilibrium is characterized by two cutoffs $\{{}^*v, v^*\}$, where ${}^*v + v^*$ in general differs from 0. The equilibrium now may be "hybrid"; for example, the majority activists may choose to be transparent while the minority activists hide in a safe space.[21] The equilibrium conditions are straightforward extensions of conditions (1)-(3), and are given in Online Appendix A.

4.1 Ideological shifts to the right or to the left

Let us investigate the marginal impact of a small change in the distribution $F(v; \rho)$, where ρ is a rotation parameter, starting from an equilibrium characterized by cutoffs $\{{}^*v, v^*\}$; for conciseness, I focus on a positional image and a safe-space equilibrium.[22] I look at the impact of a small rise in right-wing ideology. As we will see, the impact of this evolution depends on where it is located in the type distribution:

Right-wing polarization: A right-wing polarization corresponds to a shift in the distribution with $F_\rho = 0$ for $v \leq 0$ and $F_\rho \leq 0$ for $v \geq 0$.[23] A special case of right-wing polarization is a *surge in right-wing extremism.* Such a surge is characterized by $F_\rho(v) = 0$ for $v \leq v^*$ and $F_\rho(v) \leq 0$ for $v > v^*$.

Increase in acceptance of right-wing ideas: $F_\rho(v) \leq 0$ for $v \in ({}^*v, v^*)$ and $F_\rho(v) = 0$ otherwise.

[21] To see this, suppose that with probability $(1-\varepsilon)$ the type is drawn from a distribution F with support $\mathbb{R}^+$; with probability ε, the type is drawn from some distribution G with support $\mathbb{R}^-$. So, for ε small, the issue is almost consensual. Under a positional image (a safe space may emerge under the max norm when the issue is consensual. This does not occur with a positional image), right-wing activists disclose their behavior not only to their in-group, but also to everyone; in contrast (the small number of) left-wing activists hide in a safe space provided that h is not too large.

[22] The existence of a safe-space equilibrium requires that h be small enough and that the distribution not be too asymmetric.

[23] Note that a polarization (as in Section 3.2) combines a right-wing polarization with a left-wing one ($F_\rho = 0$ for $v \geq 0$ and $F_\rho \geq 0$ for $v \leq 0$).

Proposition 5 *(asymmetric distribution). Suppose a safe space equilibrium (h is small enough) and positional image concerns.*

(i) An increase in right-wing polarization boosts the right-wing safe space as well as, due to the amalgam effect, the left-wing safe space.

(ii) An increase in acceptance of right-wing ideas boosts the left-wing safe space and contracts the right-wing one.

4.2 Binary action

The binary-action model has $a \in \{0,1\}$ and $v \in \mathbb{R}^+$. It can be viewed as a very asymmetric version of the trinary-action model where F has support $\mathbb{R}^+$: $a=-1$ is irrelevant when $v \in \mathbb{R}^+$,[24] $a=0$ corresponds to being passive, and $a=1$ to being a right-wing activist.

Suppose that agents select $a_i \in \{0,1\}$, that action $a_i=0$ is a non-action (the duplicity model would give the exact same results, as noted in Section 2), and that action $a_i=1$ can at cost h be made visible only to the in-group. Let $\bar{v}=E_F[v]$ denote the prior mean of $v \in \mathbb{R}^+$. For the additive form for example:

A *safe-space equilibrium* satisfies both[25]

$$v^* - c + \int_{v^*}^{+\infty} \big[r(M^+(v^*),v) - r(M^-(v^*),v)\big]dF(v) = h \tag{4}$$

$$\int_0^{v^*} \big[r(\bar{v},v) - r(M^+(v^*),v)\big]dF(v) \geq h.$$

A *transparent equilibrium*'s cutoff satisfies[26]

$$v^* - c + \int_{0}^{+\infty} \big[r(M^+(v^*),v) - r(M^-(v^*),v)\big]dF(v) = 0 \tag{5}$$

$$\int_0^{v^*} \big[r(M^-(v^*),v) - r(M^+(v^*),v)\big]dF(v) \leq h$$

(and similarly for the mixed-strategy region).

While the characterization of equilibrium is similar to that encountered thus far in the paper, the economics are rather different from those of the symmetric case.

(i) Right-wing activists are much less fearful of being observed. Indeed, in the positional reputation case, only the transparent equilibrium exists, as $M^+(v^*)-\bar{v} = F(v^*)[M^+(v^*)-M^-(v^*)] > 0$.[27]

[24] Compared with $a=0$, action $a=-1$ goes against authenticity for an agent with $v>0$ (as $v(-1)-c<0$), and delivers the same (a worse) reputation when covert (transparent).

[25] I assume that $S(v^*) \equiv v^* - c + \int_{v^*}^{+\infty}[r(M^+(v^*),v) - r(M^-(v^*),v)]dF(v)$ is increasing in v^*.

[26] If $T(v^*) \equiv v^* - c + \int_0^{+\infty}[r(M^+(v^*),v) - r(\bar{v},v)]dF(v)$ is increasing in v^*, which I assume.

[27] And so, $\int_0^{v^*} \mu\theta(v)[\bar{v} - M^+(v^*)]dF(v) < 0 \leq h$ for all h.

(ii) In contrast, in the polar case of the maximum norm (with $v \in [0, V]$), the safe-space equilibrium obtains for the uniform distribution as long as $h \leq h_1$ for some $h_1 > 0$; for any rotation ρ of the distribution $F(v; \rho)$ around $\bar{v}$, starting from the uniform distribution, the safe-space equilibrium obtains for $h \leq h_1(\rho)$ where $h_1(\rho)$ is increasing in ρ. As we would expect, the fear of the "minority" facing the "majority" depends on the relative size and distance between the two. The safe-space equilibrium exists when opinions differ widely and $a = 0$ is prevalent; for example, with the maximum norm, hiding occurs when c is large (so v^* is also large) and differences in opinions are important (V is large).[28]

5 Extensions, applications and discussion

5.1 The dynamics of divisive behaviors

Consider the dynamic version of the bilateral-reputations, symmetric-distribution model. Time is indexed by $\tau = 0, 1, \cdots, +\infty$ and the discount factor is equal to $\delta < 1$. Each agent i sequentially selects behaviors $b_{i,0}, b_{i,1}, \ldots, b_{i,\tau}, \ldots$. For expositional simplicity, I consider two polar cases, $h = 0$ (so the equilibrium will involve safe spaces in each period) and h large (so transparency will prevail). Memory is perfect, so each agent recalls all past information received about another agent's past behaviors when assessing the latter's type. In either case, no hiding cost is incurred and agent i maximizes the present discounted value of per-period payoffs:

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

where $R_{i,\tau}$ is i's reputational payoff at the end of date τ.[29]

Interestingly, the static outcome is still an equilibrium in the safe-space case, while a repeated-action outcome under transparency follows a "Coasian pattern" and over time, puts more and more pressure on neutral types to take side.

To grasp the intuition for the *safe-space* case (h low), consider a tentative stationary equilibrium, in which each period agents play as in the static game. An active agent ($|v_i| \geq v^s$) shares his behavior with his in-group, but not with his out-group. Suppose that this active agent changes his behavior and becomes passive. His former out-group does not observe the change in behavior and thus does not infer anything. His former in-group observes that he defected and updates his reputation from $M^+(v^s)$ to some $\hat{v}$. Because

[28] Suppose that the distribution of v is uniform on $[0, V]$ and consider the possibility of a safe space equilibrium. Under the maximum norm, the reputational payoff is $-\mu V/2$ in a safe space and $-\mu(V + v^*)/2$ when not hiding. One has $v^* = c + h + (\mu V/2)$, while the condition for wanting to hide is $\mu v^*/2 \geq h$. Thus a safe space equilibrium exists iff $\mu(c + h + \mu V/2) \geq 2h$.

[29] For example, in a transparent equilibrium, in which agent i has (universal) reputation $\hat{v}_{i,\tau}$, $R_{i,\tau} \equiv \int_{-\infty}^{+\infty} r(\hat{v}_{i,\tau}, v) dF(v)$.

such behavior is off the equilibrium path for the in-group, one has some leeway in specifying beliefs, but a reasonable assumption is that $\hat{v} = v^s$ (v^s is the type in $[v^s, +\infty)$ who has the least to lose from such a deviation). Even under such a favorable updating (one could select much lower reputations), a stronger version (also satisfied by the four illustrations in Section 3) of Assumption 4(iv), namely $\int_{v^*}^{+\infty} [r(M^+(v^*), v) - r(v^*, v)] dF(v) > 0$ then implies that such a deviation reduces the agent's utility. Intuitively, the deviation to passivity does nothing to ingratiate with the out-group under safe spaces, while it is frowned upon by the in-group. A similar reasoning applies to passive players who deviate and become active, as their tardy conversion is viewed with suspicion.

Consider next *transparency* (h is very high). As the demand for reputation reflects a desire to appear moderate, the agent builds a reputation for moderation by remaining passive during a few periods. The audience then "knows" that he is not an extremist and over time becomes more and more tolerant of his activism; that is, the stigma from engaging openly in activism is time-decreasing.[30]

Proposition 6 *(dynamics under safe spaces and transparency). Suppose that agents select actions $\{a_{i,\tau}\}_{\tau \geq 0}$ sequentially and that audiences have perfect recall.*

(i) Steady behavior under safe spaces. Under safe spaces (low hiding cost), the static equilibrium ($a_i = 1$ iff $v_i \geq v^s$ where v^s is given by (4) if the solution to (4) is positive and equal to 0 otherwise) is still an equilibrium.

(ii) Progressive emboldenment under transparency. Under transparency, there exists a Coasian equilibrium in which the cutoff v^*_τ decreases over time toward $v^* = c$ (authentic behavior).

Online Appendix I illustrates the Coasian equilibrium in the transparency case.

5.2 Outings and coming outs

The very demand for safe spaces implies that one of the worst fears of a member of a community is to be outed.[31] In practice, outings tend to be more frequent for high-image-concerns members (politicians, celebrities, local notables...). While the theory developed so far predicts why such members, being more visible, are hurt more by the outing, it does

[30] This equilibrium behavior resembles that of a buyer in a bargaining or durable-good game; over time, refusals by the buyer leads to a lower and lower perception of his type by the seller and therefore a more and more accommodating stance. This accommodative stance is a lower price demand by the seller in the bargaining/durable good game, and a more moderate and thus favorable reputation in our game.

[31] I am interested in outings in a divisive-issue context. Outing of a consensual (mis-)behavior, as in the case of hypocrites (say, a politician running on family values and discreetly leading a dissolute life), of corrupt politicians, sexual abusers or people who beat up the homeless, has different welfare consequences. Such behaviors are not divisive to the extent that even their perpetuators would not claim the moral high ground for them and if push came to shove, would only invoke excuses. For consensual behaviors, "no one is so bad that he also wants to seem bad".

not explain why they are the targets of outings; to be certain, failed blackmails might be an explanation, but many outings seem to have another explanation. In line with empirical evidence that exposure to celebrities from stigmatized groups reduces prejudice (Alrababa'h et al. 2021), presumably because we know, and identify with, them, we may posit that the outing of a celebrity, successful or admired person changes the out-group's image of the community/in-group: It makes the community more mainstream, less threatening to and more like the out-group (this can be captured as a one-sided decrease in polarization as in Section 4.1). The direct implication of this assumption is that militant members of the in-group may want to out members who have a positive image in the public.

Online Appendix J develops a simple version of this argument. It further shows that outings (which lack consent) and coming outs (which by contrast are voluntary) may be complements. The outings-activated improvement of the community's image with the out-group also makes a safe space less necessary and therefore triggers coming-outs. Even if a coming out is not contemplated, an alternative motivation for outing a celebrity would be to reduce the damage caused by a fortuitous public disclosure (the safe space is not fully safe, so hiding is only probabilistic).

5.3 Endogenous social graphs and ghettoisation

A different cost of hiding from the out-group is that one may have to limit one's social graph (friends, colleagues, club mates) to agents who have similar views (as demonstrated by their behavior) and therefore will not disclose one's behavior to the out-group, either because they feel empathy, or because such disclosure may (a) reveal that one belongs to the safe space oneself, and/or (b) trigger retaliation through a similar disclosure. Members of a safe space have a common interest in respecting each other's privacy and avoiding gossiping with outsiders about their belonging to the safe space. Keeping the information private is more difficult under social mixity. I capture this in a stark form: Agent i, when choosing $a_i = +1$, benefits from a safe space (his choice of action remains hidden) if and only if his social graph is composed only of agents j such that $a_j = +1$ (and similarly for $a_i = -1$).

Reorienting one's social graph involves, first, a loss of opportunities (friends are selected in a smaller group, leading to a lower average match quality along some other dimension/personal trait) or of diversity (if diversity is valued in and of itself). Second, it involves a transition cost of making new friends if one starts with a diversified circle of friends.

To capture the first cost, associated with the *lack of diversity*, suppose that the best matching opportunities are orthogonal to the v-dimension and so reflect the overall population. Other "second-best" opportunities come at unit cost κ_δ.[32] Thus agents choosing

[32] Intuitively, friends comprise only a small proportion of the overall population. I keep a continuum for expositional simplicity, but this implies nothing as to the relative masses of friends and overall audience.

$a = 1$ must replace the $F(v^*)$ first-best friends in their out-group by $F(v^*)$ second-best friends in the in-group, at total cost of lost diversity $h(v^*) \equiv \kappa_\delta F(v^*)$. The hiding cost now grows as fewer agents act (as v^* grows). In the absence of other hiding cost, a *safe space* equilibrium[33] $v^* = v^s$ satisfies

$$S(v^s, 1) = \kappa_\delta F(v^s) \quad \text{and} \quad R_1^s(v^s, 1) - R_1^t(v^s, 1) \geq \kappa_\delta F(v^s)$$

where, as earlier, we apply the bilateral-reputations paradigm:

$$S(v^*, 1) \equiv v^* - c + \int_{v^*}^{+\infty} [r(M^+(v^*), v) - r(M^-(v^*), v)] dF(v).$$

The endogenous hiding cost interestingly is a factor of *strategic complementarity.* To see this, suppose that more agents retreat in safe spaces. Then, there is more diversity in safe spaces and the loss of diversity or the cost of matching friends with one's behavior is lower. That lowers the cost of securing privacy, making safe spaces more attractive. One can indeed find examples in which (*a*) Assumption 4(v) holds (the equilibrium is unique when the hiding cost h is exogenous), but (*b*) there are multiple equilibria when $h = \kappa_\delta F(v^*)$.

Dynamics of the social graph. As in Section 5.1, let us now index periods by $\tau \in \{0, 1 \ldots\}$. I now distinguish between the two different costs of a selective social graph: The recurring one associated with a loss of diversity/lost opportunities, and the one-shot cost associated with the effort involved in making new friends. To capture this switching cost, assume that each agent has a fixed number of friends and that there is a unit cost of changing friends κ_σ, inducing total cost of reshuffling one's social graph to peers equal to $\kappa_\sigma F(v^*)$. Conversely, recreating diversity by moving to transparency creates diversity gain $\kappa_\delta F(v^*)$, but imposes a loss again equal to $\kappa_\sigma F(v^*)$. Let h_τ denote the date-τ hiding cost. For example, the date-τ cost of abandoning a transparent behavior is

$$h_\tau \equiv (\kappa_\delta + \kappa_\sigma) F(v^*).$$

The first, diversity cost (κ_δ) is recurrent; the second switching cost (κ_σ) is transitory. Once the agent has changed friends to accommodate his privacy demand, the corresponding cost is sunk. This implies that social graphs exhibit an interesting *hysteresis*: It is costly for agents to morph their social graph toward a safe-space compatible one, but, once this is done, safe spaces will be hard to undo. Such ghettoisation may happen as religious, ethnic or linguistic communities live in good understanding, and all at once an exogenous event (killing, symbolic act, war abroad. . .) makes their identity more salient, temporarily

[33] A *transparent* equilibrium $v^* = v^t$ does not require limiting the diversity of the circle of friends ($h \equiv 0$) and satisfies

$$T(v^t, 0) = 0 \quad \text{and} \quad R_s(v^t, 0) - R_t(v^t, 0) \leq \kappa_\delta F(v^t)$$

where, as earlier:

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

increasing the intensity of image concerns μ. The mixing of the two communities may be permanently undone even after the identity turns less salient again.

Proposition 7 *(endogenous hiding costs). When the hiding cost is generated by a lack of diversity and/or a cost of switching acquaintances, the new features are: (i) Its endogeneity ($h(v^*)$, with $h'(v^*) > 0$) is a factor of strategic complementarity. (ii) The individual's social graph exhibits hysteresis.*

6 Collateral damages: From shelter to tribe

Belonging to a safe space has consequences for its members and for the broader society that go beyond those described so far. For one thing, it limits the individuals' access to a diversity of views. Levy (2021), using Facebook data, finds that consumption of ideologically congruent news on social media exacerbates polarization. For another thing, safe spaces may directly push agents to be more radical than they would be outside a safe space; this section focuses on this latter effect. So far, signaling occurred entirely through the choice of action a_i. In practice, there is often *additional signaling within the safe space.* Such "internal signaling" can explain a range of behaviors, from campus boycotts to the spreading of fake news and of conspiracy theories (railing against vaccines, "Obamagate", etc) to sheer acts of aggression against the outgroup. There are two possible rationales for this.

a) Signaling to the community. The first reason why within-in-group signaling occurs is that agents want to show they are "the true believers". They do not want to be perceived as spreaders of group-adverse messages, and conversely they will share narratives whose validity is dubious but fit the group's motivated beliefs. This implies an information that does not spread properly within the population and a weakening of democratic life and tolerance.

b) Leveraging of the fear of exclusion or outing. The extra signaling (biased narratives, actions hostile to the out-group…) considered above is voluntary, even though it may reduce social welfare and even be inefficient for the community. But the community also holds power vis-à-vis its members as it can exclude or out them. It can therefore require some actions that members would not voluntarily choose by themselves but serve the leadership or the community as a whole.

6.1 One-upmanship within the safe space

Let us illustrate the first motive, uncoerced signaling to the community (the treatment of organizational blackmail is similar). Suppose that supp $(V) = [-V, +V]$ and that $r(\hat{v}, v) \equiv \mu\theta(v)\hat{v}$ (positional reputations). Positional images are particularly favorable to internal signaling as $r_1(\hat{v}, v) > 0$ for all positive $\hat{v}$ and v. At cost c, an individual can as earlier engage in normal/minimal compliance in activity $|a_i| = 1$. He can also show zeal and pick action $A > 1$, valued vA, at cost $C > c$, still within the safe space. I assume that such zeal is dissipative in that no type would choose action A under full privacy:

$$VA - C < V - c \Longleftrightarrow V < \frac{C - c}{A - 1} \tag{6}$$

Recalling that under a positional image the agents behave authentically under transparency (h high) because $\mu E_v[\theta(v)]\hat{v} \equiv 0$ for all $\hat{v}$, this assumption also implies that in a transparent outcome no-one would take action A. Thus, zeal may arise only because agents are taking refuge in a safe space. Can a safe-space outcome with only moderate behavior (a) arise? The cutoff would then given by, letting $\Theta(v^*) \equiv \mu \int_{v^*}^{+\infty} \theta(v) dF(v)$:

$$v^* - c + \Theta(v^*)[M^+(v^*) - M^-(v^*)] = h$$

Now consider an off-path deviation to extremist action A. Applying refinement D1, such a deviation must be interpreted within the safe space as coming from type V. So no equilibrium with only moderate behavior can arise, provided that

$$\Theta(v^*)[V - M^+(v^*)] > (C - c) - V(A - 1) \tag{7}$$

If this condition is satisfied, then an equilibrium with types in $[v^*, v^{**})$ picking the moderate action and types in $[v^{**}, V]$ selecting the extremist one[34] satisfies, letting $M(v_1, v_2)$ denote the mean over interval (v_1, v_2):

$$v^* - c + \Theta(v^*)[M(v^*, v^{**}) - M^-(v^*)] = h \tag{8}$$

and[35]

$$\Theta(v^*)[M^+(v^{**}) - M(v^*, v^{**})] = (C - c) - v^{**}(A - 1). \tag{9}$$

Proposition 8 *(one-upmanship). Suppose that agents can act in a moderate* (a) *or extremist* (A) *fashion, and that zeal (the choice of A) is dissipative (condition (6)). Assume further that image is positional.*

(i) No agent adopts an extremist behavior under transparency.

(ii) In contrast, if the hiding cost is small enough that a safe space outcome prevails, extremism (and possibly only extremism) arises under condition (7) and D1.

Note that wasteful signaling here exerts negative externalities on members of the safe space; if we were to analyze overall welfare as in Section 3.3 and if furthermore the extremist action negatively impacted the out-group as in the examples given above, the cost would be even larger. The two new dark sides of safe spaces (internal and external),

[34] It may be the case that all active types choose the extremist action if the latter is not too costly. Then $v^{**} - c + \Theta(v^{**})[M^+(v^{**}) - M^-(v^{**})]] = h$, while the second equation is an inequality: $\Theta(v^{**})[M^+(v^{**}) - v^{**}] \geq (C - c) - v^{**}(A - 1)$, as D1 beliefs following off-path deviation to $a = 1$ are then $\hat{v} = v^{**}$.

[35] Note that for a given v^*, the RHS of the following equation is decreasing in v^{**}, while the LHS is increasing for a unimodal distribution F. It therefore defines an increasing function v^{**} of v^*, which can be substituted in the previous condition, yielding at most one solution v^*.

whether they arise from voluntary or coerced internal signaling, shed light on the use of “tribes” in the title of the paper. While we identified the conditions under which the creation of safe spaces have either socially beneficial effects (they can then legitimately be called “shelters”) or adversarial effects (through the externality on neutral agents, who are suspect in both communities), we establish these collateral effects of the formation of safe spaces as the very reason why such communities turn into “tribes”. Illustrations of such zeal may be wokism and the current anti-wokism. In a community of like-minded agents, a willingness to hear alternative views signals wavering, the absence of true commitment. Silo thinking is but a consequence of signaling within a safe space.[36]

Remark. The choice of an extreme action ($A > a$) is but one of the many ways an agent can signal to their in-group. Using insights on Subsection 5.3 on endogenous social graphs, an interesting alternative arises under the following circumstances: Suppose I suspect that some of my friends would be reluctant to remain my friends if they thought there is a decent chance that I support an autocrat. In order to have a good reputation with those friends, it is essential that I make them almost certain that I disapprove of the autocrat. One way of signaling this is to act transparently, thereby signaling that I really do not like the autocrat and I am not playing a double game. Technically, this can be formalized by having two distinct groups picking say $a = 1$. One group does not like the autocrat, but not so much that its members want to forgo beneficial relationships with agents who support the autocrat. The other activists are willing to make such a sacrifice and limit their social graph to the autocrat’s strong opponents (the cost of a restricted social graph is smaller for agents who are the most hostile to the autocrat, and so can serve as an effective signal). One can thereby have a schism within the activist group. This brings us to the possibility that the group can work on the prevention of schisms, studied in the next subsection.

6.2 Schisms

The escalation of partisanship studied in Proposition 8 was facilitated by the possibility for agents to over-signal without leaving the community. In turn, a community may protect itself from such escalation (assuming it can or wants to, which need not be the case) by excluding extremists, depriving the latter from an audience. To illustrate this point again in the simplest possible way, I keep assuming that the agent’s type v is distributed on $[-V, +V]$, that reputation is positional, and that the hiding cost is small so that a safe-space equilibrium prevails. As in Section 6.1, I augment the action space to include two elements on each side, say on the right side $a = 1$ and $A > 1$, so an activist can behave as a simple militant (intrinsic motivation va) or as a radical (intrinsic motivation vA).

[36] Internal signaling also implies that one must be cautious in not overestimating polarization from the group’s individual behaviors. Canen et al (2020) make a similar point in a rather different context in their work on unbundling actual polarization in Congress from changes in institutions renforcing party discipline.

Despite the dissipative nature of the extremist action (condition (6)), high-v types signal their strong convictions if (7) holds and the equilibrium is given by $\{(8),(9)\}$. Suppose now that moderates decide to create a safe space prohibiting the extremist action. Assume that extremists stay on board despite the prohibition of action A. Note that the reputation when belonging to the action-a-only safe space, jumps for $\Theta(v^*)M(v^*, v^{**})$ to the higher $\Theta(v^*)M^+(v^*)$. A deviating type choosing action A (necessarily outside this safe space) would lose material payoff from condition (6). What about the reputational payoff? Suppose action A is taken covertly (whether the deviating agent hides or not is irrelevant under positional payoffs and isolated behavior); the deviating agent loses reputation payoff $\Theta(v^*)\big[M^+(v^*) - M^-(v^*)\big] > 0$ relative to his joining the safe space prohibiting extreme behavior, and therefore faces a double whammy.

Proposition 9 *(schisms). In the one-upmanship model above, if the moderates control the group in the safe space, there exists an equilibrium in which moderates ban extremist behavior* (A) *and extremists do not secede.*

7 Related literature

That the costs and benefits of transparency hinge on individuals' reputational concerns is a central theme of the literature on prosocial behavior, in which there is broad agreement as to what represents "good" and "bad" behavior (a vast majority of people view selfishness, pollution or crime as bad and charitable contributions or public good provision as good). Lab-and-field evidence has consistently confirmed the theoretical prediction that giving a socially valued behavior more visibility makes it more prevalent.[37] The demand for reputation is then independent of the audience.[38] Ellingsen and Johannesson (2008) formalize the idea that the value of esteem depends on the source; they assume that highly moral onlookers put more weight on perceived moral traits of others. The behavior is consensual rather than divisive, and safe spaces are therefore not part of the analysis.

[37] See Ashraf-Bandiera (2018) and Bursztyn-Jensen (2017) for overviews of this literature. References include Freeman (1997), Ariely et al. (2009), Ashraf et al. (2014), Bursztyn et al. (2020) for charitable contributions, Hergueux et al. (2025) for public goods provision, Gerber et al. (2008), Funk (2010), DellaVigna et al (2017), Perez-Truglia-Cruces (2017) for voting, Ashraf et al (2014), Karing (2024) for health, and Lacetera et al. (2012) for blood donations. There is also a large experimental literature that manipulates the subjects' self-image concerns and reaches the same conclusion. Finally, good behavior in the public sphere may provide a "moral license" for bad behavior in the private one, making the result that transparency increases prosocial behavior ambiguous in a multi-tasking framework (Hong et al. 2025).

[38] Ali-Bénabou (2020) and Bénabou-Tirole (2006, 2011, 2025) are a few (of the many) illustrations of the prosocial model. Recent papers have extended our knowledge on such signaling incentives. For example, the sender can garble the performance signal (Ball 2025). She may under-consume to avoid the ratchet effect (Bonatti-Cisternas 2020). She may not dare to speak her mind if there is some correlation between the policy stance she would like to take and a socially undesirable type (Jann-Schottmüller 2020's "chilling effect"); for example, she may be afraid to speak in favor of drug liberalization by fear this might suggest drug consumption. Relatedly, she may refrain from checking into a drug rehab center or sharing info with physician if there is no assurance of privacy (Daughety-Reinganum 2010).

In the literature on conformity,[39] agents ceteris paribus want to match their actions to their intrinsic preferences, but social pressure commands them to also pick an action that mimics the average action in the population (or minimize the average distance with others). The demand for conformity reflects a societal consensus on what constitutes a desirable type (here a moderate type rather than a higher type as in the prosociality literature). This paper in contrast considers divisive issues and allows for discriminatory and endogenous visibility of one's behavior, creating scope for the emergence of safe spaces. The set of issues under investigation is accordingly different.

The analysis of behavior regarding divisive issues shares with the literature on signaling to multiple audiences[40] the idea that an agent ideally would want to change their tune depending on the audience. The signaling space (the dual choice of an action and of its disclosure) and the pattern of signaling are specific to this paper. In particular, unlike the multi-audience literature, I allow the degree of transparency to vary endogenously and formalize the notion of a safe space and its implications.

The sharing of a space with individuals with similar preferences is reminiscent of Buchanan (1965)'s theory of clubs. The emphasis of that literature however is on excludability (there is none in my paper until the section on outing) and cost sharing (my model has privately provided actions), and not on image concerns (the cornerstone of this paper).

The paper also contributes to the broader social-science debate on which of authenticity and transparency best promotes social welfare. Social scientists share the idea that people distort their public actions due to social-reputational payoffs.[41] "Authenticity" in philosophy usually has a positive connotation associated with emancipation, a view that has much influence on current laws and privacy activism. However, authenticity may well reduce social well-being if it makes us less mindful of others. A perceived anonymity on the Internet or in a big city may make us behave more in conformity with our true preferences, and yet lead to asocial behavior. The paper derives insights about how the endogeneity of the public and private spheres in our lives affects our well-being in a divisive-issue context.

[39] E.g. Bagwell-Bernheim (1996), Bernheim (1994), Corneo-Jeanne (1997), Manski-Mayshar (2003), Kuran-Sandholm (2008), Michaeli-Spiro (2015, 2017), Braghieri (2024).

[40] Austen-Smith-Fryer (2005), Bar-Isaac and Deb (2014), Bursztyn et al (2017), Frenkel (2015), Gertner et al (1988), Spiegel-Spulber (1997)

[41] For example, *"In the thought of Kant and of others influenced by him, all genuinely moral considerations rest, ultimately and at a deep level, in the agent's will. [...] To act morally is to act autonomously, not as the result of social pressure."* Bernard Williams (1985). Sartre contrasted authentic behavior ("being oneself") with actions aimed at appearing to be a certain kind of person and at conforming to established behavioral patterns to secure a more comfortable existence. Heidegger stressed that only in the private sphere can individuals be authentic, that is reveal their true self. Facing an audience, they put on a mask and build a narrative of their self. In that, the authenticity question is closely related to sociologist Erving Goffman (1956)'s "self-presentation theory", and to the literatures on "impression management" in psychology and on "image/reputation/signaling concerns" in economics.

8 Avenues for future research

Even though blind spots remain, the study of consensual issues and pro-social behavior is a well-trodden path. In contrast, social interactions in the realm of divisive issues has been a neglected field. The paper developed a conceptual framework to study such environments. When people do not agree on what's right or wrong and hostile opinions weigh more than favorable ones, transparency leads agents to either alter their behavior or take refuge in a safe space. The paper applied the framework to show that, as envisioned by privacy advocates, safe spaces act as shelters against value destruction (discrimination, violence...). But they also have dark sides as they involve internalized costs (reduced use of public spaces or diversity of social graph, duplicity), create reputational externalities on moderates (who are suspected by both sides and pushed to pick one), and generate tribalistic over-signaling beyond desired practice (either voluntarily, or coerced through the threat of outing) that would not occur under transparency. We also saw how symmetric increases in polarization lead to more retreat into safe spaces, while the impact of right-wing polarization depends on whether it comes from an increase in the extremist population's size or from a greater acceptance of right-wing ideas in the non-activist population. We then showed that safe spaces are dynamically stable and even subject to hysteresis.

The concept of a "safe space" (a place where individuals' views can be fully expressed without fear of violence, harassment, or hate speech) has occasionally been unduly extended to include protection against different opinions, as the section on one-upmanship suggests would happen. Social norms within a like-minded group may create a surveillance society that has nothing to envy that developed by autocratic governments. As many have noted a safe space should not stifle freedom of speech.[42] This quest for an authenticity enabled by a greater freedom to assume one's identity may be socially beneficial, as in the case of anonymity in mental health fora. But it may also induce a "ghettoisation of thinking", and a reduced tolerance for debate. As embodied in my framework, the contours of the private and public spheres are not set only by technology, they are also socially determined by explicit individual choices.

What issues are considered divisive is country- and epoch- specific. Research should try to understand the drivers of this evolution: Technological progress (which may generate new controversies, as in the case of medically assisted reproduction, or, as in the case of social networks, may create new safe spaces and at the same time magnify the impact of outings); geopolitical tensions and wars (reinforcing identities); economic factors (affecting the size of social graphs or altering the importance of parental inputs in child education); importance of religion; etc. Pluralistic ignorance also affects divisiveness, and so does its dispelling.

[42] Political correctness (Morris 2001), like safe spaces, has been a welcome evolution, but may be abused by those who refuse dialogue and tolerance vis-à-vis others who don't think like them and want a space that is expunged of individuals with conflicting opinions (Lukianoff-Haidt 2018).

In this paper, individuals manipulate their public image through acts and disclosure decisions. But others may also take the individual's public image in a direction that the latter would not wish as shown by the rise of doxing, facilitated by technology and social networks and employed for various purposes, from culture wars to cyber-criminality. The defining feature of doxing is the enlistment of popular justice to damage an individual's public image through the disclosure of unpopular attitudes or embarrassing personal traits, and possibly the sharing of the person's address, phone number, social security number and so on. While ignoring doxing, my model contains the rationale for it: A malicious intent to discourage others from expressing their difference. I leave this and other extensions to future work.

References

Adriani, F., and S. Sonderegger (2019), "A Theory of Esteem Based Peer Pressure," *Games and Economic Behavior*, 115: 314–335.

Ali, S.N., and R. Bénabou (2020), "Image Versus Information: Changing Societal Norms and Optimal Privacy," *American Economic Journal: Microeconomics*, 12(3): 116–164.

Alrababa'h, A., Marble, W., Mousa, S., and A. Siegel (2021), "Can Exposure to Celebrities Reduce Prejudice? The Effect of Mohamed Salah on Islamophobic Behaviors and Attitudes," *American Political Science Review*, 115(4): 1–18.

An, Mark Y. (1998), "Logconcavity versus Logconvexity: A Complete Characterization," *Journal of Economic Theory*, 80(2): 350–69.

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

Ashraf, N. and O. Bandiera (2018), "Social Incentives in Organizations?" *Annual Review of Economics*, 10: 439–463.

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

Austen-Smith, D., and R. Fryer (2005), "An Economic Analysis of 'Acting White"' *Quarterly Journal of Economics,* 120(2): 551–583.

Bagwell, L. and D, Bernheim (1996), "Veblen Effects in a Theory of Conspicuous Consumption," *American Economic Review*, 86(3): 349–373.

Ball, I. (2025), "Scoring Strategic Agents," *American Economic Journal: Microeconomics*, 17(1): 97–129.

Bar-Isaac, H., and J. Deb (2014), "(Good and Bad) Reputation for a Servant of Two Masters," *American Economic Journal: Microeconomics*, 6(4): 293–325.

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

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

Bénabou, R., and J. Tirole (2025), "Laws and Norms," *Journal of Political Economy*, forthcoming.

Bernheim, D. (1994), "A Theory of Conformity," *Journal of Political Economy*, 102: 841–77.

Bonatti, A. and G. Cisternas (2020), "Consumer Scores and Price Discrimination," *Review of Economic Studies*, 87(2): 750–791.

Bouvard, M., and R. Levy (2017), "Two-Sided Reputation in Certification Markets," *Management Science*, 64: 4755–4774.

Braghieri, L. (2024), "Political Correctness, Social Image, and Information Transmission," *American Economic Review*, 114(2): 3877–3904.

Buchanan, J. (1965), "An Economic Theory of Clubs," *Economica*, 32(125): 1–14.

Bursztyn, L., and R. Jensen (2017), "Social Image and Economic Behavior in the Field: Identifying, Understanding and Shaping Social Pressure," *Annual Review of Economics*, 9: 131–153.

Bursztyn, L., Fujiwara, T., and A. Pallais (2017), "'Acting Wife': Marriage Market Incentives and Labor Market Investments," *American Economic Review*, 107(11): 3288–3319.

Bursztyn, L., Gonzalez, A., and D. Yanagizawa-Drott (2020), "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia," *American Economic Review*, 110(10): 2997–3029.

Bursztyn, L., Haaland, I., Rao, A. and C. Roth (2020), "I Have Nothing Against Them, But...," mimeo.

Canen, N., C. Kendall, and F. Trebbi (2020), "Unbundling Polarization," *Econometrica*, 88(3): 1197–1233.

Corneo, G., and O. Jeanne (1997), "Conspicuous Consumption, Snobbism, and Conformism," *Journal of Public Economics*, 66(1): 55–71.

Daughety, A., and J. Reinganum (2010), "Public Goods, Social Pressure, and the Choice between Privacy and Publicity," *American Economic Journal: Microeconomics*, 2(2): 191–221.

DellaVigna, S., List, J., Malmendier, U. and G. Rao (2017), "Voting to Tell Others," *Review of Economic Studies*, 84: 143–181.

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

Freeman, R. (1997), "Working for Nothing: The Supply of Volunteer Labor," *Journal of Labor Economics*, 15(1): S140–66.

Frenkel, S. (2015), "Repeated Interaction and Rating Inflation: A Model of Double Reputation," *American Economic Journal: Microeconomics*, 7(1): 250–280.

Fromm, E. (1941), *Escape from Freedom*, Farrar & Rinehart.

Funk, P. (2010), "Social Incentives and Voter Turnout: Evidence from the Swiss Mail Ballot System," *Journal of the European Economic Association*, 8(5): 1077–1103.

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

Gertner, R., Gibbons, R. and D. Scharfstein (1988), "Simultaneous Signalling to the Capital and Product Markets," *Rand Journal of Economics*, 19(2): 173–190.

Goffman, E. (1956), *The Presentation of Self in Everyday Life*, Open Library.

Henderson, R. and E. McCready (2024), *Signaling without Saying. The Semantics and Pragmatics of Dogwhistles*, Oxford University Press.

Hergueux, J., Algan, Y., Benkler, Y., and M. Fuster-Morell (2025), "Public Good Superstars: A Lab-in-the-Field Study of Wikipedia," *Economic Journal*, 135(667): 861–891.

Hong, F., Tirole, J. and C. Zhang (2025), "Moral Licensing: Prosocial Behavior in Public and Private Spheres," mimeo.

Jann, O., and C. Schottmüller (2020), "An Informational Theory of Privacy," *Economic Journal*, 130: 93–124.

Jewitt, I. (2004), “Notes on the ‘Shape’ of Distributions,” unpublished.

Karing, A. (2024), “Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone,” *Quarterly Journal of Economics*, 139(4): 2083–2133.

Kuran, T. and W. Sandholm (2008), “Cultural Integration and its Discontents,” *Review of Economic Studies*, 75(1): 201–228.

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

Levy, R. (2021), “Social Media, News Consumption, and Polarization: Evidence from a Field Experiment,” *American Economic Review*, 111(3): 831–870.

Lukianoff, G., and J. Haidt (2018), *The Coddling of the American Mind*, Penguin Books.

Manski, C., and J. Mayshar (2003), “Private Incentives and Social Interactions: Fertility Puzzles in Israel,” *Journal of the European Economic Association*, 1(1): 181–211.

Michaeli, M. and D. Spiro (2015), “Norm Conformity across Societies,” *Journal of Public Economics*, 132: 51–65.

Michaeli, M. and D. Spiro (2017), “From Peer Pressure to Biased Norms,” *American Economic Journal: Microeconomics*, 9(1): 152–216.

Morris, S. (2001), “Political Correctness,” *Journal of Political Economy*, 109: 231–265.

Perez-Truglia, R. and G. Cruces (2017), “Partisan Interactions: Evidence from a Field Experiment in the United States,” *Journal of Political Economy*, 125(4): 1208–1243.

Simmel, G. (1906), “The Sociology of Secrecy and Secret Societies,” *American Journal of Sociology*, 11(4): 441–498.

Spiegel, Y., and D. Spulber (1997), “Capital Structure with Countervailing Incentives,” *Rand Journal of Economics*, 28(1): 1–24.

Williams, B. (1985), *Ethics and the Limits of Philosophy*, Harvard University Press.

Safe Spaces: Shelters or Tribes?

Jean Tirole

Online Appendix

A Bilateral reputations: Equilibrium existence and uniqueness

Demand for reputation.

I here consider an agent's demand for reputation vis-à-vis groups of agents choosing $a = -1$, $a = 0$, and $a = 1$, respectively.

Lemma A1 *(benefit from appearing as moderate under transparency). Under Assumption 4, the reputational benefit of an agent with homogeneous reputation $\hat{v}$ (as is the case under transparency), $\int_{-\infty}^{+\infty} r(\hat{v}, v)dF(v)$, is symmetric and concave in reputation $\hat{v}$ and peaks at $\hat{v} = 0$. It is strictly concave whenever $r_{11} < 0$ (by contrast it is flat at 0 in the positional image case).*

Proof of Lemma A1. We need to show that $\int_{-\beta}^{+\beta} r(\hat{v}, v)dF(v)$, which is concave from Assumption 4(iii), peaks at 0. Its derivative at 0 is

$$\int_{-\beta}^{+\beta} r_1(0, v)dF(v) = \int_{0}^{+\beta} r_1(0, v)dF(v) + \int_{-\beta}^{0} r_1(0, v)dF(v).$$

Assumption 4(i) implies that $r_1(\hat{v}, v) = -r_1(-\hat{v}, -v)$ and so $r_1(0, -v) = -r_1(0, v)$, implying that $\int_{-\beta}^{+\beta} r_1(0, v)dF(v) = 0$. ■

Consider a symmetric equilibrium with cutoffs $-v^*$ and v^*, and conduct the thought-experiment in which active agent i selects whom to disclose his action $|a_i| = 1$ to, assuming that they observe the others' action (although not their type). This thought-experiment corresponds to the case in which disclosure strategies have no direct impact on payoffs ($h \equiv 0$).

Lemma A2 *(demand for reputation). Consider a symmetric equilibrium. Under Assumption 4, and ignoring any cost of self-presentation, an agent i who selects $|a_i| = 1$ strictly prefers to disclose his behavior to his peers, and prefers not to disclose his behavior to non-peers (strictly so unless $v^* = 0$ and $x = 0$); and so $x_i = 1$.*

Proof of Lemma A2: We focus on the behavior of agents $v \geq v^*$ (by symmetry, this also determines the behavior of agents $v \leq -v^*$).

Consider first the agent's peers. When disclosing $a_i = 1$ to them, agent i's image with these peers is $M^+(v^*) \equiv E[v|v \geq v^*]$. When not disclosing, the image vis-à-vis the peers is

$$\hat{v} = M_x^-(v^*) \equiv \frac{xF(-v^*)}{xF(-v^*) + [F(v^*) - F(-v^*)]} M^-(-v^*),$$

since $E[v| - v^* \leq v \leq v^*] = 0$ where $M^-(v^*) \equiv M_1^-(v^*) = E[v|v \leq v^*]$. We need to show that

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

Note that $\hat{v} \leq 0 \leq v^*$. So, from part (ii) of Assumption 4, a sufficient condition for this inequality to hold is

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

which is guaranteed by part (iv) of Assumption 4.

Next, consider the disclosure of $a_i = 1$ to the passive group ($v \in [-v^*, v^*]$).[43] Not disclosing yields reputation $\hat{v} = 0$ while disclosing leads to image $\hat{v} = M^+(v^*) > v^*$. Lemma A1 implies that the gain from non-disclosure is non-negative (strictly positive if $r_{11} < 0$):

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

Finally, consider the disclosure of $a_i = 1$ to the group of agents choosing $a_i = -1$ ($v \leq -v^*$). In the absence of disclosure, the latter attribute reputation

$$\hat{v} = M_x^+(-v^*) \equiv \frac{x[1 - F(v^*)]}{x[1 - F(v^*)] + [F(v^*) - F(-v^*)]} M^+(v^*) = -M_x^-(v^*),$$

and so the gain from non-disclosure is:

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

from Assumption 4(ii). ■

[43] In a symmetric equilibrium, a passive agent j who does not observe agent i's choice formulates posterior distribution $F(v; x)$ on the latter's type, where:

$$F(v; x) \equiv \begin{cases} \dfrac{xF(v)}{1 - 2(1 - x)F(-v^*)} & \text{for } v \leq -v^* \\ \dfrac{xF(-v^*) + [F(v) - F(-v^*)]}{1 - 2(1 - x)F(-v^*)} & \text{for } v \in [-v^*, v^*] \\ \dfrac{xF(-v^*) + [F(v^*) - F(-v^*)] + x[F(v) - F(v^*)]}{1 - 2(1 - x)F(-v^*)} & \text{for } v \geq v^*. \end{cases}$$

These posterior beliefs are not well-defined when $x = v^* = 0$. Then $a = 0$ is an off-equilibrium-path action (negative types pick $a = -1$, positive types $a = +1$, and $v = 0$ is indifferent between the two but prefers them to $a_i = 0$). As discussed in Section 2.1, we will then naturally assume that posterior beliefs put all weight on $v = 0$: $F(v; 0) = 0$ for $v < 0$, $= 1$ for $v > 0$.

To sum up, an active agent ceteris paribus wants to share the nature of his behavior with his peers, but not with his non-peers. This will indeed be the case (and so $x = 1$) if hiding is costless.

The baseline reputational payoffs applied to bilateral reputations write:

$$\begin{cases} R_s(v^*, x) & \equiv \displaystyle\int_{v^*}^{+\infty} r(M^+(v^*), v)dF(v) + \int_{-v^*}^{v^*} r(0, v)dF(v) + \int_{-\infty}^{-v^*} r(M_x^+(-v^*), v)dF(v) \\ R_t(v^*) & = \displaystyle\int_{-\infty}^{+\infty} r(M^+(v^*), v)dF(v) \\ R_\varnothing(v^*, x) & \equiv \displaystyle\int_{v^*}^{+\infty} r(M_x^-(v^*), v)dF(v) + \int_{-v^*}^{v^*} r(0, v)dF(v) + \int_{-\infty}^{-v^*} r(M_x^+(-v^*), v)dF(v). \end{cases}$$

Assumption 4 delivers some useful properties. Note that for all $\{v^* \neq 0,\, x \neq 0\}$,[44]

$$R_s(v^*, x) > \max\{R_t(v^*), R_\varnothing(v^*, x)\}, \tag{A.1}$$

where $R_s > R_t$ results from Lemma A2 and $R_s > R_\varnothing$ from part (iv) of Assumption 4 (since $R_s(v^*, x) - R_\varnothing(v^*, x) \equiv \int_{v^*}^{+\infty} [r(M^+(v^*), v) - r(M_x^-(v^*), v)dF(v)] > 0$). Similarly, for all $v^* \geq 0$,

$$R_\varnothing(v^*, 0) \geq R_t(v^*)$$

with strict inequality if $r_{11} < 0$.[45] Note also that $\partial M_x^-/\partial x < 0$ and $r_1 > 0$ for $v > \max\{0, \hat{v}\}$ imply that

$$\frac{\partial}{\partial x}(R_s(v^*, x) - R_\varnothing(v^*, x)) > 0.$$

In words, an increase in the use of safe spaces (in x) creates more suspicion on passive agents relatively to active agents who hide and raises the incentive to not disclose when selecting $|a_i| = 1$: There are *strategic complementarity in hiding*.[46]

[44] For $v^* = 0$, $x > 0$, $R_t(0, x) = R_t(0)$.

[45] Part (i) of Assumption 4 implies that

$$\begin{aligned} R_\varnothing(v^*, 0) - R_t(v^*) &= \int_{-\infty}^{+\infty} [r(0, v) - r(M^+(v^*), v)]dF(v) \\ &= \int_{0}^{+\infty} [2r(0, v) - r(M^+(v^*), v) - r(-M^+(v^*), v)]dF(v). \end{aligned}$$

Part (iii) of Assumption 4 implies that for all v

$$r(0, v) \geq \frac{r(M^+(v^*), v) + r(-M^+(v^*), v)}{2}$$

with strict inequality if r is strictly concave in $\hat{v}$.

[46] The externality on passive agents is captured in:

$$\frac{\partial R_\varnothing}{\partial v^*} = 2f(v^*)[r(0, v^*) - r(M_x^-(v^*), v^*)] + 2\left[\int_{v^*}^{+\infty} r_1(M_x^-(v^*), v)dF(v)\right] \frac{dM_x^-(v^*)}{dv^*} > 0$$

Both terms on the RHS of this equation are strictly positive if $x > 0$ (and both are equal to 0 if $x = 0$, since $M_0^-(v^*) \equiv 0$ for all v^*). The first term corresponds to the extensive margin: The marginal contributor

Similarly, the increased suspicion on passive players as x increases implies that

$$\frac{\partial}{\partial x}(R_t(v^*) - R_{\varnothing}(v^*, x)) > 0.$$

Proposition A1 *Assume that μ is not too large, so as to ensure uniqueness of the cutoffs. Under Assumption 4, there exists a unique equilibrium and it is symmetric.*

- *(i) There exist h_1 and h_2, with $h_1 < h_2$, such that the equilibrium is a safe-space equilibrium ($x = 1$) if and only if $h < h_1$, and a transparent equilibrium ($x = 0$) if and only if $h > h_2$. The equilibrium is in mixed strategy over (h_1, h_2), with x decreasing continuously with h over that range.*

- *(ii) As the hiding cost increases, then the threshold v^s in a safe space equilibrium increases. There is more activity than under full privacy (the authentic self level), i.e. $v^s \leq c$, in a safe-space equilibrium for $h = 0$, and less activity than under full privacy in a transparent equilibrium ($v^t \geq c$, with a strict inequality if the reputational payoff is strictly concave in reputation: $r_{11} < 0$).*

The proof of this proposition follows directly from the following property. Under Assumption 4(i) and (iii), for any $\beta > 0$, $\int_{-\beta}^{+\beta} r(\hat{v}, v)dF(v)$, which is concave in $\hat{v}$ (strictly so if $r_{11} < 0$), peaks at $\hat{v} = 0$.

Proof of non-existence of asymmetric equilibria

Incentive compatibility requires that there exist ${}^*v$ and v^*, with ${}^*v \leq v^*$ such that $a_i = +1$ if $v_i > v^*$, $a_i = -1$ if $v_i < {}^*v$ and $a_i = 0$ if ${}^*v < v < v^*$. Let ${}^*x$ and x^* denote the probabilities of hiding in a safe space when picking actions -1 and $+1$, respectively. Let

$$\begin{aligned} M^+_{x,v^*}({}^*v) \equiv\; & \frac{x[1-F(v^*)]}{[F(v^*) - F({}^*v)] + x[1-F(v^*)]} M^+(v^*) \\ & + \frac{F(v^*) - F({}^*v)}{[F(v^*) - F({}^*v)] + x[1-F(v^*)]} M({}^*v, v^*) \end{aligned}$$

(recall that $M({}^*v, v^*)$ is the mean over the interval $[{}^*v, v^*]$).

Let

$$M^-_{x,-v^*}(-{}^*v) \equiv -M^+_{x,v^*}({}^*v).$$

We repeatedly use the identity:

$$\int_{-\infty}^{V} r(\hat{v}, v)dF(v) \equiv \int_{-V}^{+\infty} r(-\hat{v}, v)dF(v).$$

for all V and $\hat{v}$.

has a more tolerant image of a passive player when he is himself passive (he is less suspicious). The second term corresponds to the inframarginal active agents; when v^* increases, the conditional mean for $v < v^*$ increases, implying more tolerance.

We need to generalize Assumption 4(v) to the asymmetric-behavior case in order to guarantee uniqueness of the payoffs. Let

$$L(v^*, {}^*v) \equiv v^* - c + \int_{v^*}^{+\infty} [r(M^+(v^*), v) - r(M^-_{1,{}^*v}(v^*), v)] dF(v).$$

Assumption 5 *For all* ${}^*v \leq v^*$,

$$\begin{aligned} L(v^*, {}^*v) &= L(-{}^*v, -v^*) \ \Rightarrow \ v^* = -{}^*v \\ L(v^*, {}^*v) &> L(-{}^*v, -v^*) \ \Rightarrow \ v^* > -{}^*v. \end{aligned}$$

Consider first a *transparent equilibrium*, and let $M({}^*v, v^*)$ denote the mean conditional on $v \in [{}^*v, v^*]$. Then

$$\begin{aligned} v^* - c + \int_{-\infty}^{+\infty} r(M^+(v^*), v) dF(v) &= \int_{-\infty}^{+\infty} r(M({}^*v, v^*), v) dF(v) \\ &= -{}^*v - c + \int_{-\infty}^{+\infty} r(M^-({}^*v), v) dF(v). \end{aligned}$$

And so

$$v^* + \int_{-\infty}^{+\infty} r(M^+(v^*), v) dF(v) = -{}^*v + \int_{-\infty}^{+\infty} r(M^+(-{}^*v), v) dF(v).$$

Assumption 5 then implies that $v^* = -v^*$, and so any transparent equilibrium must be symmetric.

Next suppose that both cutoff types are indifferent between hiding and not hiding. Then, because the reputational gain when choosing $a_i = 1$ and hiding rather than choosing $a_i = 0$ purports only to the in-group, for v^* we have:

$$v^* - c + \int_{v^*}^{+\infty} [r(M^+(v^*), v) - r(M^-_{{}^*x,{}^*v}(v^*), v)] dF(v) = h.$$

For ${}^*v$ we have:

$$\begin{aligned} &-{}^*v - c + \int_{-\infty}^{{}^*v} [r(M^-({}^*v), v) - r(M^+_{x^*,v^*}({}^*v), v)] dF(v) \\ &= -{}^*v - c + \int_{-{}^*v}^{+\infty} [r(M^+(-{}^*v), v) - r(-M^-_{x^*,-v^*}(-{}^*v), v)] dF(v) = h. \end{aligned}$$

When ${}^*x = x^* = 1$, Assumption 5 then implies that $v^* = -v^*$. Otherwise, without loss of generality, ${}^*x$ belongs to $(0, 1)$ and x^* belongs to $(0, 1]$. Let $\hat{v}_0$ denote the beliefs of passive agents. For $x^* \in (0, 1]$ we have:

$$\begin{aligned} &-\int_{-\infty}^{+\infty} r(M^+(v^*), v) dF(v) + \int_{-\infty}^{{}^*v} r(M^+_{x^*,v^*}({}^*v), v) dF(v) + \int_{{}^*v}^{v^*} r(\hat{v}_0, v) dF(v) \\ &\quad + \int_{v^*}^{+\infty} r(M^+(v^*), v) dF(v) \geq h, \end{aligned}$$

with equality when $x^* \neq 1$. For ${}^*x \in (0,1)$ we have:

$$\begin{aligned} h = &-\int_{-\infty}^{+\infty} r(M^-({}^*v), v)dF(v) + \int_{-\infty}^{{}^*v} r(M^-({}^*v), v)dF(v) + \int_{{}^*v}^{v^*} r(\hat{v}_0, v)dF(v) \\ &+ \int_{v^*}^{+\infty} r(M^-_{{}^*x,{}^*v}(v^*), v)dF(v) = -\int_{-\infty}^{+\infty} r(M^+(-{}^*v), v)dF(v) \\ &+ \int_{-{}^*v}^{+\infty} r(M^+(-{}^*v), v)dF(v) + \int_{{}^*v}^{v^*} r(\hat{v}_0, v)dF(v) + \int_{-\infty}^{-v^*} r(M^+_{{}^*x,-{}^*v}(-v^*), v)dF(v), \end{aligned}$$

where the last equality is due to Assumption 4(iv). Now, if we add the difference between the equations for v^* and x^* and the difference between the equations for ${}^*v$ and ${}^*x$, we have:

$$v^* + \int_{-\infty}^{+\infty} r(M^+(v^*), v)dF(v) \leq -{}^*v + \int_{-\infty}^{+\infty} r(M^+(-{}^*v), v)dF(v),$$

with equality when $x^* \neq 1$. Assumption 4(v) again implies that $v^* = -{}^*v$ when $x^* \neq 1$, and $v^* \leq -v^*$ if $x^* = 1$. If $x^* \neq 1$, besides $v^* = -{}^*v$, with a simple algebra we have:

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

Therefore, Assumption 4(ii) implies that $x^* = {}^*x$.

Hence, we only need to check the case in which $x^* = 1$ and ${}^*x$ belongs to $(0,1)$. Assumption 4(ii), and the fact $M^-_{{}^*x,{}^*v}(v^*) > M^-_{1,{}^*v}(v^*)$, imply that:

$$\int_{v^*}^{+\infty} r(M^-_{{}^*x,{}^*v}(v^*), v)dF(v) > \int_{v^*}^{+\infty} r(M^-_{1,{}^*v}(v^*), v)dF(v).$$

Using the equation for v^* and ${}^*v$, we have:

$$\begin{aligned} &v^* - c + \int_{v^*}^{+\infty} [r(M^+(v^*), v) - r(M^-_{1,{}^*v}(v^*), v)]dF(v) - h > 0 \\ &\quad = -{}^*v - c + \int_{-{}^*v}^{+\infty} [r(M^+(-{}^*v), v) - r(M^-_{1,-v^*}(-{}^*v), v)]dF(v) - h. \end{aligned}$$

Assumption 5 implies that $v^* > -v^*$, a contradiction.

The last case to be studied is when $x^* = 0$ and ${}^*x$ belongs to $(0,1]$. The equation for v^* is:

$$\begin{aligned} &v^* - c + \int_{-\infty}^{+\infty} r(M^+(v^*), v)dF(v) \\ &\quad = \int_{{}^*v}^{+\infty} r(M^-_{{}^*x,{}^*v}(v^*), v)dF(v) + \int_{-\infty}^{{}^*v} r(M({}^*v, v^*), v)dF(v). \end{aligned}$$

When the set of agents picking $a = 1$ is a transparent group ($x^* = 0$), then:

$$\begin{aligned} &\int_{-\infty}^{+\infty} r(M^+(v^*), v)dF(v) + h \\ &\geq \int_{v^*}^{+\infty} r(M^+(v^*), v)dF(v) + \int_{{}^*v}^{v^*} r(M^-_{{}^*x,{}^*v}(v^*), v)dF(v) + \int_{-\infty}^{{}^*v} r(M({}^*v, v^*), v)dF(v). \end{aligned}$$

The difference between these two equations yields:

$$v^* - c + \int_{v^*}^{+\infty} [r(M^+(v^*), v) - r(M^-_{{}^*x,{}^*v}(v^*), v)]dF(v) - h \leq 0$$

The symmetry of the distribution of types and Assumption 4(i) entail that the equation giving ${}^*v$ is:

$$\begin{aligned} -{}^*v - c + \int_{-{}^*v}^{+\infty} r(M^+(-{}^*v), v)dF(v) - h \\ = \int_{-{}^*v}^{+\infty} r(M(-v^*, -{}^*v), v)dF(v). \end{aligned}$$

Assumption 4(ii), and the fact that $M(-v^*, -{}^*v) > M^-(-{}^*v)$, imply that:

$$\begin{aligned} -{}^*v - c + \int_{-{}^*v}^{+\infty} [r(M^+(-{}^*v), v) - r(M^-_{1,-v^*}(-{}^*v), v)]dF(v) - h > 0 \\ \geq v^* - c + \int_{v^*}^{+\infty} [r(M^+(v^*), v) - r(M^-_{{}^*x,{}^*v}(v^*), v)]dF(v) - h \end{aligned}$$

Hence Assumption 5 implies that $v^* < -v^*$, when ${}^*x = 1$. The equation for ${}^*x$ is:

$$\begin{aligned} \int_{-\infty}^{v^*} r(M^-({}^*v), v)dF(v) + \int_{{}^*v}^{+\infty} r(M^-_{{}^*x,{}^*v}(v^*), v)dF(v) - h \\ \geq \int_{-\infty}^{v^*} r(M^-({}^*v), v)dF(v) + \int_{{}^*v}^{+\infty} r(M^-({}^*v), v)dF(v), \end{aligned}$$

using Assumption 4(i), and the symmetry of the distribution of types yields:

$$\begin{aligned} \int_{-\infty}^{-{}^*v} r(M^+_{{}^*x,-{}^*v}(-v^*), v)dF(v) - h \\ \geq \int_{-\infty}^{-{}^*v} r(M^+(-{}^*v), v)dF(v), \end{aligned}$$

with equality when ${}^*x \neq 1$. Combining equations for ${}^*x$, ${}^*v$, and v^*, we have:

$$\begin{aligned} v^* - c + \int_{-\infty}^{+\infty} r(M^+(v^*), v)dF(v) \\ \geq -{}^*v - c + \int_{-\infty}^{+\infty} r(M^+(-{}^*v), v)dF(v), \end{aligned}$$

with equality when ${}^*x \neq 1$. Again Assumption 5 implies $v^* \geq -v^*$. But we know that $v^* < -v^*$ when ${}^*x = 1$, a contradiction. Also, if ${}^*x \neq 1$, besides $v^* = -{}^*v$ (from Assumption 5), combining equations for x^*, ${}^*v$, and v^* yields:

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

Therefore, Assumption 4(ii) implies that ${}^*x \leq 0$, a contradiction.

■

Constructive proof of equilibrium existence under a positional image

Let

$$\begin{cases} R_s(v^*,x) \text{ denote the image payoff when choosing } a_i=1 \text{ and hiding it from non-peers} \\ R_t(v^*) \text{ denote the image payoff when choosing } a_i=1 \text{ and being transparent} \\ R_\varnothing(v^*,x) \text{ denote the image payoff when choosing } a_i=0. \end{cases}$$

Let $\Theta(v^*) \equiv \mu \int_{v^*}^{+\infty} \theta(v)dF(v)$. Suppose that individuals who act (say, $a_i=1$) hide with probability x and remain transparent with probability $1-x$.

Then

$$\begin{cases} R_s(v^*,x) & = \Theta(v^*)M^+(v^*) - \Theta(v^*)M^+(v^*)\dfrac{x[1-F(v^*)]}{x[1-F(v^*)]+[2F(v^*)-1]} \\ R_t(v^*) & = 0 \\ R_\varnothing(v^*,x) & = -2\Theta(v^*)M^+(v^*)\dfrac{x[1-F(v^*)]}{x[1-F(v^*)]+[2F(v^*)-1]} \end{cases}$$

Using $M^+(-v^*) = -M^-(v^*) = \frac{1-F(v^*)}{F(v^*)}M^+(v^*)$, the mixed-strategy region is then characterized by the following conditions

$$v^*-c+R_s(v^*,x)-R_\varnothing(v^*,x)=h \;\Leftrightarrow\; v^*-c+\Theta(v^*)\left[\frac{2x[1-F(v^*)]+2F(v^*)-1}{x[1-F(v^*)]+2F(v^*)-1}\right]M^+(v^*)=h \tag{A.2}$$

$$R_s(v^*,x)-R_t(v^*)=h \;\Leftrightarrow\; v^*-c=-2\Theta(v^*)\left[\frac{x[1-F(v^*)]}{x[1-F(v^*)]+[2F(v^*)-1]}\right]M^+(v^*) \tag{A.3}$$

The redundant condition implied by (A.2) and (A.3), is type $v \geq v^*$'s indifference between transparency and safe space when $a=1$:

$$\Theta(v^*)\left[\frac{2F(v^*)-1}{x[1-F(v^*)]+[2F(v^*)-1]}\right]M^+(v^*)=h. \tag{A.4}$$

To prove existence of an equilibrium in mixed strategy, let, for an arbitrary cutoff v,

$$T(v,x) \equiv v-c+2\Theta(v)\frac{x[1-F(v)]}{x[1-F(v)]+[2F(v)-1]}M^+(v)$$

denote the cutoff type's net gain of choosing $a=1$ and being transparent rather than choosing $a_i=0$, and thereby avoiding the two-sided suspicion that arises when $a_i=0$; and let

$$S(v,x) \equiv v-c+\Theta(v)\left[\frac{2x[1-F(v)]+[2F(v)-1]}{x[1-F(v)]+[2F(v)-1]}\right]M^+(v)$$

denote the gross gain of picking $a_i=1$ and hiding (this gain ignores the hiding cost h) relative to picking $a_i=0$.

Note that both T and S are strictly increasing in x. Furthermore, the suboptimality of transparency can be rewritten as $T(v^*, 1) \leq 0$ and that condition (A.5), given that $\Delta(v^*) = M^+(v^*)/F(v^*)$, amounts to $S(v^*, 1) = h$.

To guarantee the existence of an interior solution ($v^* > 0$), let us assume that $S(0, 1) < 0$, or

Assumption 6 *(interior solution).* $c > 2\Theta(0)M^+(0)$.

Conditions (A.2) and (A.3) are equivalent to $S(v^*, x) = h$ and $T(v^*, x) = 0$, respectively.

Next, for $v < c$, we can define the function

$$x(v) \equiv \frac{[2F(v) - 1](c - v)}{[2\Theta(v)M^+(v) + (v - c)][1 - F(v)]}$$

so that $T(v, x(v)) = 0$.

Note that

$$x(v) > 0 \quad \Leftrightarrow \quad 2\Theta(v)M^+(v) + v - c > 0.$$

Because $2\Theta(0)M^+(0) - c < 0$ and $2\Theta(c)M^+(c) > 0$, there exists an interval $[b, c]$ such that $0 < b < c$,

$$2\Theta(v)M^+(v) + b - c = 0.$$

And so

$$2\Theta(v)M^+(v) + v - c > 0 \quad \text{for } v \in (b, c].$$

Restricting attention to the interval $(b, c]$, straightforward computations show that

$$S(v, x(y)) = \frac{2\Theta(v)M^+(v) + v - c}{2}$$

and so $x(v) > 0 \Leftrightarrow S(v, x(v)) > 0$.

Now define the function $y(v)$ on $(b, c]$ by

$$y(v) = \min\{x(v), 1\}.$$

And let $y(b) \equiv 1$ (as $\lim_{v \to b^+} x(v) = +\infty$).

So let $Z(v) \equiv S(v, y(v))$, defined on $[b, c]$. This function is continuous and satisfies:

$$Z(v) = S(v, y(v)) < S(v, x(v)) = 0 \ \text{ for } v \text{ close to } b$$

and

$$Z(c) = \Theta(c)M^+(c) > 0.$$

Define $Z(b)$ as $S(v,1)$.

The mean-value theorem implies that for all $h \in [Z(b), T(c))$ there exists v^* such that

$$S(v^*, y(v^*)) = Z(v^*) = h,$$

and

$$Z(v^*, y(v^*)) = \begin{cases} 0 & \text{if} \quad y(v^*) < 1 \\ \leq 0 & \text{if} \quad y(v^*) = 0. \end{cases}$$

This proves the existence of a mixed-strategy equilibrium. ■

B Total image payoff and welfare

(i) The total image payoffs under full privacy, safe spaces and transparency are:

$$\begin{aligned} \mathcal{R}^{fp} &= \int_{-\infty}^{+\infty} r(0,v)dF(v) \\ \mathcal{R}^{s} &= \int_{-v^*}^{v^*} r(0,v)dF(v) + 2[1-F(v^*)]\left[\int_{v^*}^{+\infty} r(M^-(v^*),v)dF(v)\right] \\ &\quad +2[1-F(v^*)]^2\left[\int_{v^*}^{+\infty} [r(M^+(v^*),v) - r(M^-(v^*),v)]dF(v)\right] \\ \mathcal{R}^{t} &= [2F(v^*)-1]\left[\int_{-\infty}^{+\infty} r(0,v)dF(v)\right] + 2[1-F(v^*)]\left[\int_{-\infty}^{+\infty} r(M^+(v),v)dF(v)\right]. \end{aligned}$$

And so

$$\mathcal{R}^{fp} - \mathcal{R}^{t} = 2[1-F(v^*)]\left[\int_{-\infty}^{+\infty} [r(0,v) - r(M^+(v^*),v)]dF(v)\right] \geq 0$$

from Lemma 2 (with strict inequality unless $r_{11} = 0$). Next,

$$\mathcal{R}^{fp} - \mathcal{R}^{s} = 2[1-F(v^*)]\left[\int_{v^*}^{+\infty} \big[r(0,v) - F(v^*)r(M^-(v^*),v) - [1-F(v^*)]r(M^+(v^*),v)\big]dF(v)\right].$$

Recall that r is concave in $\hat{v}$ and that for all v^*,

$$F(v^*)M^-(v^*) + [1-F(v^*)]M^+(v^*) = 0$$

from the martingale property. And so, for all v

$$r(0,v) \geq F(v^*)r(M^-(v^*),v) + [1-F(v^*)]r(M^+(v^*),v).$$

Alternative proof. Let B denote the total reputational payoff of others vis-à-vis audience type v. For example, under full transparency $B^{ft}(v) \equiv \int_{-\infty}^{+\infty} r(\tilde{v},v)dF(\tilde{v})$. Along these lines, the total reputational payoffs under full privacy, safe space, mixed and transparent equilibria are, when $v \geq v^*$,

$$\begin{aligned}
B^{fp}(v) &\equiv r(0,v)\\
B^{ss}(v^*,v) &\equiv [1-F(v^*)]r(M^+(v^*),v)+F(v^*)r(M^-(v^*),v)\\
B^{m}(v^*,x,v) &\equiv [1-F(v^*)]r(M^+(v^*),v)+(1-x)F(-v^*)r(M^-(-v^*),v)\\
&\quad +[2F(v^*)-1+xF(-v^*)]r(M_x^-(v^*),v)\\
B^{t}(v^*,v) &\equiv [1-F(v^*)]r(M^+(v^*),v)+F(-v^*)r(M^-(-v^*),v)\\
&\quad +[2F(v^*)-1]r(0,v).
\end{aligned}$$

For each $v \geq v^*$, type v's information structures is such the distributions of the conditional means are ordered mean-preserving spreads. Concavity ($r_{11} \leq 0$) then implies that for all $v \geq v^*$ and for given $\{v^*, x\}$

$$B^{fp}(v) \geq B^{ss}(v^*,v) \geq B^m(v^*,x,v) \geq B^t(v^*,v) \geq B^{ft}(v).$$

For example, to compare the three possible equilibrium configurations, it suffices to demonstrate that, for $x > y$, then $B^m(v^*,x,v) \geq B^m(v^*,y,v)$. To show this, note that

$$B^m(v^*,x,v) \geq B^m(v^*,y,v) \iff \alpha r(M_x^-(v^*),v) \geq \beta r(M^-(-v^*),v)+\gamma r(M_y^-(v^*),v)$$

where $\alpha \equiv [2F(v^*)-1]+xF(-v^*)$, $\beta \equiv (x-y)F(-v^*)$, and $\gamma \equiv [2F(v^*)-1+yF(-v^*)]$ and so $\alpha = \beta + \gamma$. The martingale property, $\alpha M_x^-(v^*) \equiv \beta M^-(-v^*) + \gamma M_y^-(v^*)$, yields the result.

A similar reasoning applies to an audience type $v \in [-v^*, v^*]$ and (by sheer symmetry) to $v \leq -v^*$. Finally, aggregating over all audience types v yields part (i) of Proposition 4.

(ii) Recall that $W(v^*,x) = R_\varnothing(v^*,x) + 2\int_{v^*}^{+\infty}(v-v^*)dF(v)$, regardless of the privacy regime. The non-image term is maximized for $v^* = c$, which is the case for full privacy, or for a positional image under transparency. As for the image term,

$$\mathcal{R}^{fp} = \int_{-\infty}^{+\infty} r(0,v)dF(v) = \mathcal{R}^t(v^*) \geq \max\{\mathcal{R}^s(v^*), \mathcal{R}^m(v^*,x)\}.$$

To demonstrate the latter inequality, let $\hat{v}_0(v)$ denote the image of a passive agent with audience v. Then, whatever the regime

$$R_0 = \int_{-\infty}^{+\infty} r(\hat{v}_0(v),v)dF(v).$$

Furthermore for $v < 0$ (resp. > 0), $\hat{v}_0(v) \geq 0$ (resp. ≤ 0), and strictly so unless $x = 0$.

$$\mathcal{R}^{fp} = \int_{-\infty}^{+\infty} r(0,v)dF(v) \geq \int_{-\infty}^{+\infty} r(\hat{v}_0(v),v)dF(v).$$

■

C Computation of equilibrium and welfare in our illustrations

(a) Positional image

Let $r(\hat{v}, v) = \mu\theta(v)\hat{v}$, where θ is antisymmetric. Let $\Delta(v^*) \equiv M^+(v^*) - M^-(v^*) = M^+(v^*)/F(v^*)$. Because F is unimodal with mode 0, the function Δ is decreasing for $v^* < 0$ and increasing for $v^* > 0$ (Jewitt 2004). Letting

$$\Theta(v^*) \equiv \mu \int_{v^*}^{+\infty} \theta(v) dF(v) \geq 0,$$

denote the intensity of image concerns vis-à-vis types $v_j \geq v^*$ under a positional image,[47]

$$\begin{cases} R_s(v^*, x) & = & \Theta(v^*)[M^+(v^*) - M_x^+(-v^*)] \\ R_\varnothing(v^*, x) & = & -2\Theta(v^*)M_x^+(-v^*) \\ R_t(v^*) & = & 0. \end{cases}$$

Let us derive the equilibrium. Technical details are provided in the Appendix.

Safe space equilibrium ($x = 1$). Letting $\Delta(v^*) \equiv M^+(v^*) - M^-(v^*)$, such an equilibrium exists if and only if for some cutoff v^*

$$v^* - c + \Theta(v^*)\Delta(v^*) = h \tag{A.5}$$

and

$$[2F(v^*) - 1]\Theta(v^*)\Delta(v^*) \geq h. \tag{A.6}$$

Note that the cutoff affects the image concerns in two opposite ways. When more agents act (v^s decreases), $\Delta(v^s)$ decreases from Jewitt's lemma (participation becomes less elitist and a lower glory within the in-group is attached to it, promoting strategic substitutability), but $\Theta(v^s)$ increases (a higher number of like-minded agents observe $a_i = 1$, promoting strategic complementarity).[48]

Importance of social approval. In the positional image model, social approval is more important under function $\tilde{\theta}$ than under function θ if $\tilde{\theta}(v) \geq \theta(v)$ for all $v \geq 0$ (so by symmetry $\tilde{\theta}(v) \leq \theta(v)$ for all $v \leq 0$). [It is also more important if μ increases.] With respect to this criterion, comparative statics with respect to the importance of social approval are straightforward: *An increase in the importance of social approval increases $\Theta(v^s)$ and leaves $\Delta(v^s)$ constant, and so v^s decreases.*

Transparent equilibrium ($x = 0$). Suppose now that agent i's behavior is observed by all

[47] We verify that $R_s(v^*, x)$ is an increasing function of x, from $\Theta(v^*)M^+(v^*)$ for $x = 0$ to $\Theta(v^*)[M^+(v^*) - M^-(v^*)] > \Theta(v^*)M^+(v^*)$ for $x = 1$.

[48] Note that $\partial(\Theta(v)M^+(v))/\partial v|_{v=0} > 0$.

$(x = 0)$.[49] The weight on each image is $\Theta(-\infty) = 0$, as any behavior creates as many supporters as opponents with the same intensity of (dis)approval.

Thus $v^* = v^t = c$ (where "t" stands for "transparency"). Let $h_2 \equiv \Theta(c)M^+(c)$. A transparent equilibrium obtains iff $h \geq h_2$.

Mixed-strategy equilibrium $(0 \leq x \leq 1)$. Such an equilibrium satisfies both $v^* - c + R_s - R_\varnothing - h = 0$ and $R_s - R_t = h$. For convenience, let us assume that $\Theta(v)M^+(v)$ is weakly increasing in v. Then (using (A.6)), $h_2 = \Theta(c)M^+(c) \geq \Theta(v^s(h_1))M^+(v^s(h_1)) = h_1 \frac{F(v^s(h_1))}{2F(v^s(h_1))-1} > h_1$. The equilibrium is depicted in Figure 1 in the text.

Welfare. Agent welfare in a safe space or mixed equilibrium, $W^{s,m}$, can be written as

$$W^{s,m} = -2\Theta(v^*)M_x^+(-v^*) + 2\int_{v^*}^{+\infty}(v - v^*)dF(v).$$

Welfare under transparency is

$$W^t = 2\int_c^{\infty}(v - c)dF(v).$$

And so, $W^t > W^{s,m}$. Transparency yields the social optimum W^{fp}, as it promotes authenticity and involves no hiding cost.

When image is zero-sum, $2[1 - F(v^*)]R_1^s(v^*, x) + [2F(v^*) - 1]R_0(v^*, x) = 0$; and so, for $x > 0$, there is too much belonging to safe spaces:

$$\frac{\partial W}{\partial v^*} = 2f(v^*)[R_1^s(v^*, x) - R_0(v^*, x)] > 0.$$

(b) Maximum norm

In a *safe space equilibrium* under the maximum norm:

$$S(v^s, 1) = v^s - c - \mu[V + M^+(-v^*)] + \mu[V + M^+(-v^*)] = 0 \iff v^s = c + h.$$

The safe space equilibrium exists as long as

$$h \leq \mu[M^+(c + h) - M^+(-c - h)].$$

Assume that $h - \mu[M^+(c + h) - M^+(-c - h)]$ is increasing in h, which is indeed the case if image concerns are not too large. Then a safe space equilibrium exists if and only if $h \leq h_1$ where $h_1 = \mu[M^+(c + h_1) - M^+(-c - h_1)]$.

In a *transparent equilibrium*, the cutoff v^t is given by

$$v^t - c - \mu[V + M^+(v^t)] + \mu V = 0 \qquad \iff \qquad v^t = c + \mu M^+(v^t).$$

[49] Incentive compatibility again implies the existence of cutoffs $-v^*$ and v^*. That means that, for all j, $a_i = +1$ creates image $\hat{v}_{ji} = M^+(v^*)$, $a_i = -1$ image $\hat{v}_{ji} = M^-(-v^*)$ and $a_i = 0$ image $\hat{v}_{ji} = M(-v^*, v^*)$, where $M(-v^*, v^*)$ is the mean in the interval $(-v^*, v^*)$, namely 0 under a symmetric distribution.

From our assumption of a monotone hazard rate for F, $0 < (M^+)' < 1$ and so $v^t < V$ if and only if $\mu < (V-c)/V$. Welfare under transparency is:

$$W^t = -\mu V + 2\int_{v^t}^{V} (v - v^t) dF(v).$$

A transparent equilibrium requires that

$$h \geq \mu M^+(v^t) \equiv h_2 > h_1.$$

The *mixed region* satisfies $v^* = v^m = c + h$, and

$$h = \mu[M^+(c+h) - M_x^+(-c-h)]$$

Assuming again that image concerns are not too large so that $1-\mu[(M^+)'(v^*)+(M_x^+)'(-v^*)] > 0$, then the equilibrium probability of hiding x is decreasing in h, with $x = 1$ for $h = h_1$ and $x = 0$ for $h = h_2$.

Overall, the equilibrium is unique and its pattern follows the general one –safe spaces, then mixed, then transparent– as h increases. The difference is that authentic behavior occurs in the safe space region rather that in the transparency one for a positional image.

Welfare in the safe space and mixed regions is

$$W^{s,m} = -\mu[V + M_x^+(-c-h)] + 2\int_{c+h}^{V} [v - (c+h)] dF(v).$$

Again, for image concerns that are not too large, $dW^{s,m}/dh < 0$.

D Proof that the modified L^p norm satisfies Assumption 4

We actually prove a stronger form of Assumption 4(iv) The individual prefers to be perceived as a representative member of the in-group than as the cutoff type v^*. This stronger form of Assumption 4(iv) in the case of the modified L^p norm (i.e. without the homogeneity condition) requires that

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

or (omitting the intensity μ of image concerns)

$$A \equiv (M^+(v^*) - v^*)^p \int_{v^*}^{+\infty} \left[\left(\frac{v - v^*}{M^+(v^*) - v^*} \right)^p - \left(\left| 1 - \frac{v - v^*}{M^+(v^*) - v^*} \right| \right)^p \right] dF(v) > 0.$$

Let $X \equiv \frac{v - v^*}{M^+(v^*) - v^*} \geq 0$ for $v \geq v^*$.

$$\begin{cases} \text{When } X \geq 1, \; X^p = ((X-1)+1)^p > (X-1)^p + p(X-1) \\ \text{When } 0 \leq X \leq 1, \; X^p \geq 0 \geq (1-X)^p - (1-X). \end{cases}$$

And so, in both cases (i.e. whenever $X \geq 0$),

$$X^p - (|1-X|)^p \geq X - 1 \text{ (with strict inequality unless } X = 0).$$

Thus

$$A > (M^+(v^*) - v^*)^p \int_{v^*}^{+\infty} [v - M^+(v^*)] dF(v) = 0. \qquad \blacksquare$$

E The non-additive case

We allow for broader reputational concerns, in particular to accommodate the (true) L^p norm (and as a special case the maximum norm). Consider two actions in $\{-1, 0, 1\}$: b for the onlooker and a for the agent. Let $d \in \{ND, D\}$ (no disclosure, disclosure) denote the agent's disclosure decision for $|a| = 1$. For a passive agent ($a = 0$) who does not have a disclosure decision, we use the convention that $d = ND$. We assume that the agent's reputational payoff is a weakly increasing function of his reputations vis-à-vis members of subgroups J_{-1}, J_0, J_{+1}:

$$R_i(v^*, x) \equiv \mu\Phi(R^d_{-1,a_i}(v^*, x) + R^d_{0,a_i}(v^*, x) + R^d_{1,a_i}(v^*, x)), \tag{A.7}$$

where $R^d_{b,a}$ is differentiable in (v^*, x).

For example for the L^p norm, $\Phi(X) \equiv X^{1/p}$ and $R^d_{1,1} = -\int_{M^+(v^*)}^{+\infty} |v - M^+(v^*)|^p dF(v)$, etc. Note that $R^D_{b,a}$ is always independent of x, while $R^{ND}_{b,a}$ is independent of a.

We make assumptions that were proved to hold for bilateral reputations:

Assumption 7 *For all (v^*, x):*

(i) Disclosing to the in-group raises the reputational payoff: $R^D_{a,a} \geq R^{ND}_{a,a}$.

(ii) Hiding from the out-group raises the reputational payoff: $R^{ND}_{b,a} \geq R^D_{b,a}$ for $b \neq a$.

(iii) The incentives to disclose to the in-group and to the out-group are increasing in x: $\frac{\partial}{\partial x}(R^D_{b,a} - R^{ND}_{b,a}) > 0$ for all b and for $a \in \{-1, +1\}$.

Let

$$S(v^*, x) \equiv v^* - c + \mu\big[\Phi(R^{ND}_{-1,1}, R^{ND}_{0,1}, R^D_{1,1}) - \Phi(R^{ND}_{-1,0}, R^{ND}_{0,0}, R^{ND}_{1,0})\big]$$

denote the cutoff's net benefit from acting in a safe space relative to being passive, and

$$T(v^*, x) \equiv v^* - c + \mu\big[\Phi(R^D_{-1,1}, R^D_{0,1}, R^D_{1,1}) - \Phi(R^{ND}_{-1,0}, R^{ND}_{0,0}, R^{ND}_{1,0})\big]$$

denote the cutoff's net benefit from acting transparently relative to being passive.

Lemma A3 *Suppose that reputational payoffs are given by the L^p norm (examples 3 and 4). Then, there exists $\bar{\mu} > 0$ such that for all $\mu \leq \bar{\mu}$, the functions $S(v^*, x)$ and $T(v^*, x)$ are strictly increasing in v^* for all x.*

Proof of Lemma A3

Let us show that Assumption 4(iv) holds for the true L^p norm if image concerns are "not too high". Formally, there is a $\bar{\mu}$ such that for all $\mu < \bar{\mu}$, functions $S(v^*, x)$ and $T(v^*, x)$ are strictly increasing in v^* for all x. Define

$$\begin{cases} K_T(v^*, x) = -\dfrac{1}{\mu}\big[R_t(v^*) - R_\varnothing(v^*, x)\big] \\ \\ K_S(v^*, x) = -\dfrac{1}{\mu}\big[R_s(v^*, x) - R_\varnothing(v^*, x)\big], \end{cases}$$

we have to show $|\frac{\partial K_i}{\partial v^*}| < M$ for some fixed M and for $i \in \{T, S\}$.

$|\frac{\partial K_i}{\partial v^*}|$ is a continuous function on any set $[0, V]$ and is therefore bounded. It thus suffices to show that there exist V and M such that $|\frac{\partial K_i}{\partial v^*}| < M$ for $v^* > V$.

We start with $R_1^t(v^*)$, and show that $0 < \partial\big(\frac{\frac{-1}{\mu}R_1^t(v^*)}{\partial v^*}\big) < 1$. This actually will always hold.

$$R_t(v^*) = -\mu\left[\int_{-\infty}^{M^+(v^*)}(M^+(v^*) - v)^p dF(v) + \int_{M^+(v^*)}^{+\infty}(v - M^+(v^*))^p dF(v)\right]^{\frac{1}{p}}.$$

Let:

$$L \equiv \frac{\Big[\int_{-\infty}^{M^+(v^*)}(M^+(v^*) - v)^{p-1}dF(v) - \int_{M^+(v^*)}^{+\infty}(v - M^+(v^*))^{p-1}dF(v)\Big]}{\Big[\int_{-\infty}^{M^+(v^*)}(M^+(v^*) - v)^p dF(v) + \int_{M^+(v^*)}^{+\infty}(v - M^+(v^*))^p dF(v)\Big]^{\frac{p-1}{p}}},$$

$$\Rightarrow \frac{\partial R_1^t(v^*)}{\partial v^*} = -\mu(M^+(v^*))' L,$$

where the hazard rate condition implies that $0 < (M^+(v^*))' < 1$. We can show L is positive:

$$\int_{-\infty}^{M^+(v^*)}(M^+(v^*) - v)^{p-1}dF(v) > \int_{-\infty}^{-M^+(v^*)}(M^+(v^*) - v)^{p-1}dF(v)$$
$$= \int_{M^+(v^*)}^{+\infty}(v + M^+(v^*))^{p-1}dF(v) > \int_{M^+(v^*)}^{+\infty}(v - M^+(v^*))^{p-1}dF(v)$$

L is also lower than 1:

$$L < \frac{\int_{-\infty}^{+\infty}|M^+(v^*) - v|^{p-1}dF(v)}{\Big(\int_{-\infty}^{+\infty}|M^+(v^*) - v|^p dF(v)\Big)^{\frac{p-1}{p}}} \leq 1,$$

which the last inequality is a special case of Hölder's inequality for a probability space and random variable X:

$$E(|X|^r) \leq (E(|X|^s))^{r/s} \qquad 0 < r < s,$$

Next, we want show that there exist V and M such that $|\frac{-1}{\mu}\frac{\partial R_0(v^*,x)}{\partial v^*}| < M$, for all $v^* > V$ and all x.

$$-\frac{1}{\mu}R_\varnothing(v^*,x) \;=\; \Big[2\big(\int_{v^*}^{+\infty}(-M_x^-(v^*)+v)^p dF(v) + \int_0^{v^*} v^p dF(v)\big)\Big]^{\frac{1}{p}}.$$

Define N_1, N_2, and D in the following expression:

$$\frac{\partial(-\frac{1}{\mu}R_\varnothing(v^*,x))}{\partial v^*} \equiv \frac{N_1+N_2}{D}$$
$$= \frac{\overbrace{\frac{2}{p}f(v^*)(v^{*p}-(v^*-M_x^-(v^*))^p +}^{N_1}\overbrace{2(-M_x^-(v^*))'\int_{v^*}^{+\infty}(-M_x^-(v^*)+v)^{p-1}dF(v)}^{N_2}}{\underbrace{\Big[2\big(\int_{v^*}^{+\infty}(-M_x^-(v^*)+v)^p dF(v) + \int_0^{v^*} v^p dF(v)\big)\Big]^{\frac{p-1}{p}}}_{D}}.$$

We now show $\frac{|N_1|}{D}$, and $\frac{|N_2|}{D}$ are bounded for all v^* and x. Let $y = v - M_x^-(v^*)$.

$$\begin{aligned}\frac{|N_2|}{D} \;&\le\; \frac{|N_2|}{2^{\frac{p-1}{p}}\big[\int_{v^*}^{+\infty}(-M_x^-(v^*)+v)^p dF(v)\big]^{\frac{p-1}{p}}} = \frac{|2(-M_x^-(v^*))'|E(y^{p-1})(1-F(v^*))}{2^{\frac{p-1}{p}}(E(y^p))^{\frac{p-1}{p}}(1-F(v^*))^{\frac{p-1}{p}}}\\ &= \frac{|2^{\frac{1}{p}}(-M_x^-(v^*))'|(1-F(v^*))^{1/p}E(y^{p-1})}{(E(y^p))^{\frac{p-1}{p}}}\\ &\le |2^{\frac{1}{p}}(-M_x^-(v^*))'|(1-F(v^*))^{1/p}\end{aligned}$$

by Hölder's inequality. Therefore $\frac{|N_2|}{D}$ is bounded for all v^* and x.

$$\frac{|N_1|}{D} = \frac{2}{p}\frac{f(v^*)v^{*p}\big(1-\big(1-\frac{M_x^-(v^*)}{v^*}\big)^p\big)}{D}.$$

We know $f(v^*)v^{*p}$ is bounded and that $\lim_{v^*\to+\infty}\dfrac{M_x^-(v^*)}{v^*} = 0$. Also D^{-1} is bounded since $D > (2\int_0^V v^p dF(v))^{1-1/p}$. Hence $\frac{|N_1|}{D}$ is bounded for all x and all $v^* > V$.

Finally we need to prove $|\frac{-1}{\mu}\frac{\partial R_s(v^*,x)}{\partial v^*}| < M$ for all $x \in [0,1]$ and $v^* > V$.

$$\begin{aligned}-\frac{1}{\mu}R_s(v^*,x) \;&=\; \Big[\int_{v^*}^{M^+(v^*)}(M^+(v^*)-v)^p dF(v) + \int_{M^+(v^*)}^{+\infty}(v-M^+(v^*))^p dF(v)\\ &+\;\; 2\int_0^{v^*} v^p dF(v) + \int_{-\infty}^{-v^*}(M_x^+(-v^*)-v)^p dF(v)\Big]^{\frac{1}{p}}.\end{aligned}$$

Define N, N_1, N_2, and D in the following way

$$\frac{\partial(-\frac{1}{\mu}R_s(v^*,x))}{\partial v^*} \equiv \frac{N}{D},$$

$$\begin{aligned} N &= \frac{1}{p}\Big[-(M^+(v^*)-v^*)^p f(v^*) + p(M^+(v^*))' \int_{v^*}^{M^+(v^*)} (M^+(v^*)-v)^{p-1} dF(v) \\ &\quad +p(-M^+(v^*))' \int_{M^+(v^*)}^{+\infty} (v-M^+(v^*))^{p-1} dF(v) + 2v^{*p} f(v^*) \\ &\quad -f(v^*)(-M_x^-(v^*)+v^*)^p + p(-M_x^-(v^*))' \int_{v^*}^{+\infty} (v-M_x^-(v^*))^{p-1} dF(v)\Big] \end{aligned}$$

$$\begin{aligned} &= \overbrace{\frac{1}{p} f(v^*) v^{*p}\Big(2-\Big(\frac{M^+(v^*)}{v^*}-1\Big)^p - \Big(\frac{-M_x^-(v^*)}{v^*}+1\Big)^p\Big)}^{N_1} \\ &+ \overbrace{M^+(v^*)' \int_{v^*}^{M^+(v^*)} (M^+(v^*)-v)^{p-1} dF(v)}^{N_{21}} \\ &+ \overbrace{(-M^-(v^*)') \int_{M^+(v^*)}^{+\infty} (V-M^+(v^*))^{p-1} dF(v)}^{N_{22}} \\ &+ \overbrace{(-M_x^-(v^*))' \int_{(v^*)}^{+\infty} (V-M_x^-(v^*))^{p-1} dF(v)}^{N_{23}} = N_1 + N_2. \end{aligned}$$

$$\begin{aligned} D &= \Big[\int_{v^*}^{M^+(v^*)} (M^+(v^*)-v)^p dF(v) + \int_{M^+(v^*)}^{+\infty} (v-M^+(v^*))^p dF(v) \\ &\quad +2\int_0^{v^*} v^p dF(v) + \int_{-\infty}^{-v^*} (M_x^+(-v^*)-v)^p dF(v)\Big]^{1-1/p}. \end{aligned}$$

Note that D is positive and D^{-1} is bounded.

$$\frac{N_2}{D} \le \frac{|N_{21}|}{D} + \frac{|N_{22}|}{D} + \frac{|N_{23}|}{D}$$

$$\begin{aligned} \frac{|N_{21}|}{D} &= \frac{|M^+(v^*)'|(F(M^+(v^*)) - F(v^*))E(y^{p-1})}{[F(M^+(v^*)) - F(v^*)]^{\frac{p-1}{p}} E(y^p)^{\frac{p-1}{p}}} \\ &= \frac{|M^+(v^*)'|[F(M^+(v^*)) - F(v^*)]^{1/p} E(y^{p-1})}{E(y^p)^{\frac{p-1}{p}}} \\ &\le |M^+(v^*)'|[F(M^+(v^*)) - F(v^*)]^{1/p}, \end{aligned}$$

by Hölder's inequality where $y = M^+(v^*) - v$. Hence $\frac{|N_{21}|}{D}$ is bounded for all v^*, x.

$$\begin{aligned} \frac{|N_{22}|}{D} &\le \frac{|(-M^+(v^*)')|(1-F(M^+(v^*))^{1/p} E(y^{p-1})}{E(y^p)^{\frac{p-1}{p}}} \\ \frac{|N_{23}|}{D} &\le \frac{|(-M_x^-(v^*))'|(1-F(v^*))^{1/p} E(y^{p-1})}{E(y^p)^{\frac{p-1}{p}}}. \end{aligned}$$

Similarly $\frac{|N_{22}|}{D}$ and $\frac{|N_{23}|}{D}$ are bounded.

$$|N_1| = \frac{1}{p} v^{*p} f(v^*) \Big| 2 - \Big(\frac{M^+(v^*)}{v^*} - 1\Big)^p - \Big(\frac{-M_x^-(v^*)}{v^*} + 1\Big)^P \Big|.$$

We know $v^{*p} f(v^*)$ is bounded and that $\lim_{v^* \to +\infty} \dfrac{M_x^-(v^*)}{v^*} = 0$. The proof is complete if we show $\frac{M^+(v^*)}{v^*}$ is bounded.

$$(M^+(v^*))' < 1 \Leftrightarrow \frac{f(v^*)}{1 - F(v^*)}[M^+(v^*) - v^*] < 1$$

$$\Leftrightarrow \quad M^+(v^*) - v^* < \frac{1 - F(v^*)}{f(v^*)} < \frac{1/2}{f(0)},$$

where the last inequality stems from monotone hazard rate property. Thus

$$0 < \frac{M^+(v^*)}{v^*} - 1 < \frac{1}{2v^* f(0)}$$

Therefore $\frac{M^+(v^*)}{v^*}$ is bounded for $v^* > V$. ■

Lemma A3, shows that for image concerns that are not too important, (the counterpart of) Assumption 4(v) is satisfied for the L^p norm (while it is not satisfied in general for the modified L^p norm, which is not homogenous, unless the support of F is finite).

Proposition A2 *(non-additive case). Suppose that an agent's overall reputational payoff is given by (A.7), where Φ is an increasing function. Under Assumptions 1-3 and 7, the characterization and existence of transparent, mixed and safe-space equilibria carries over to non-additive reputation payoffs.*

Proposition A2 allows us to extend the analysis to the L^p norm. The missing elements of its proof follow the step of the proof of Proposition A1.

F Ancillary benefits (Section 3.4)

Let us prove the claim made in Section 3.4. Consider a symmetric equilibrium configuration with cutoff v^* and active agents retreating in safe spaces (which requires h not being too large). As explained in the text, an agent engages in a second activity besides the choice of a: The agent may match (randomly) within one of safe spaces that emerge in equilibrium (for this, the agent must have joined the safe space) or within the "general pool" that potentially admits those who have revealed nothing to members of this general pool (by being either active within a safe space or passive). Passive agents necessarily match within the general pool, while active agents who have concealed their behavior to the out-group can choose to match either in their safe space ("endogamous matching") or in the general pool ("exogamous matching")[50]. Let $\mathcal{P}$ denote one of these three pools, labeled $\mathcal{P}_{-1}$, $\mathcal{P}_{\varnothing}$, $\mathcal{P}_{+1}$. To an agent i's payoff function, we add a random-matching benefit that is decreasing with distance and indexed by parameter $m > 0$; up to a constant, this benefit is (up to a constant):

$$mE_{v_j \in \mathcal{P}}[- \mid v_j - v_i \mid]$$

(a) Endogamous matching. Suppose, first, that all agents picking action $a = +1$ choose to match within $\mathcal{P} = \mathcal{P}_{+1}$. The necessary (and sufficient) condition for this is that the average distance be smaller under endogamous than under exogamous matching:

$$M^+(v^*) - v^* \leq v^* - 0.$$

Indeed, the cutoff type is at distance $M^+(v^*) - v^*$ of the average in-group member given that $\mathcal{P}_{+1} = [v^*, +\infty)$, and at distance v^* of the average participant in pool $P_{\varnothing} = (-v^*, +v^*)$. Taking the support of F to be $[-V$,+V] (where V can be arbitrarily large), the cutoff is then given by

$$v^* - c + \int_{v^*}^{V} [r(M^+(v^*), v) - r(M^-(v^*), v)]dF(v) + m[2v^* - M^+(v^*)] = h. \qquad \text{(A.8)}$$

Note that the LHS of this equation is a fortiori increasing in v^* under Assumption 4(v), provided that the monotone hazard rate condition is satisfied (then $d(2v^* - M^+(v^*))/dv \in (1,2)$). This configuration obtains, say, when c is sufficiently large (using the fact that v^* is increasing in c, and that $M^+(v^*) - v^*$ tends to 0 as v^* tends to V).

(b) Exogamous matching. When v^* decreases (say, because c does so) and reaches the level at which $M^+(v^*) = 2v^*$, the safe space matching benefit vanishes, the two choices (of a and $\mathcal{P}$) decouple, and the cutoff is given by condition (4):

$$v^* - c + \int_{v^*}^{V} [r(M^+(v^*), v) - r(M^-(v^*), v)]dF(v) = h. \qquad \text{(A.9)}$$

Pool $\mathcal{P}_{\varnothing}$ is then given by $\mathcal{P}_{\varnothing} = (-v^\dagger, +v^\dagger)$, where $v^\dagger > v^*$ is (uniquely, under a monotone hazard rate) defined by

$$M^+(v^\dagger) = 2v^\dagger.$$

Note that the LHS of (A.9) is equal to $-c$ for $v^* = 0$. To see that exogamous matching may happen in equilibrium (for $c < c^\sharp$ for some $c^\sharp > 0$), suppose that c and h are equal to 0 (or small). Then condition (A.9) yields corner solution $v^* = 0$ and so $v^\dagger > v^*$.

G Reputation as a random member of a group

The paper assumed that the representative member of his perceived group defines an agent's reputation. That is, the reputational payoff of an agent i vis-à-vis an agent j with

[50] If the cutoff in equilibrium matches in the general pool, matching is mostly, but not entirely exogamous: The match may occur with another moderate of the same safe space.

type v, when agent j attributes conditional distribution $F(\tilde{v}|v)$ with support $\mathbb{R}$ to agent i's type, is $r(E_{F(\cdot|v)}[\tilde{v}], v)$. Alternatively, we could have assumed that agent i is viewed as a random, rather than representative member of his perceived group. Then, agent i's reputational payoff with agent j is

$$\int_{-\infty}^{+\infty} r(\tilde{v}, v) dF(\tilde{v}|v).$$

The two formulations coincide for a positional image ($r(\tilde{v}, v) = \mu\theta(v)\tilde{v}$), but they differ more generally. Indeed, the law of iterated expectations implies that reputations as members of perceived groups generate a constant-sum game even when r does not satisfy the linearity assumption of the positional image case. Intuitively, animosity can be deflected/redirected, but not reduced in aggregate. We keep making Assumption 4.

Proposition A3 *(random member of group). Suppose that the reputational payoff of an agent vis-à-vis another agent of type v is $\int_{-\infty}^{+\infty} r(\tilde{v}, v)dF(\tilde{v}|v)$, where $F(\tilde{v}|v)$ is the distribution of the former agent's type conditional on the latter agent's information about his action. Then, reputation acquisition is a constant-sum game. The characterization is the same as that of the positional-image model when reputation is anchored on the type of the representative member of the group: A safe space (mixed, transparent) equilibrium exists if and only if $h \leq h_1$ (resp. $h \in [h_1, h_2]$, $h \geq h_2$) for some $h_2 > h_1 > 0$.*

We define the counterpart assumption to Assumption 4(iv) for a reputation as a random member of a group:

Assumption 4(iv)′ *(benefit from being perceived by the in-group as representative of the in-group rather than as a passive type).*

$$\int_{-\infty}^{-v^*} \Big(\int_{-\infty}^{-v^*} \frac{r(\tilde{v}, v)}{1 - F(v^*)} dF(\tilde{v}) \Big) dF(v) \geq \int_{-\infty}^{-v^*} \Big(\int_{-v^*}^{+v^*} \frac{r(\tilde{v}, v)}{2F(v^*) - 1} dF(\tilde{v}) \Big) dF(v).$$

Assumption 4(iv)′ is satisfied for a positional image.

Consider a symmetric equilibrium $\{v^*, x\}$. And let

$$\Gamma^1(v, v^*) \equiv \frac{\int_{v^*}^{+\infty} r(\tilde{v}, v)dF(\tilde{v})}{1 - F(v^*)} \equiv \Gamma^{-1}(-v, -v^*),$$

where $\Gamma^{-1}(v, -v^*) \equiv \dfrac{\int_{-\infty}^{-v^*} r(\tilde{v}, v)dF(\tilde{v})}{F(-v^*)}$, and $\Gamma^0(v, v^*) \equiv \dfrac{\int_{-v^*}^{+v^*} r(\tilde{v}, v)dF(\tilde{v})}{2F(v^*) - 1}$,

denote the reputational payoff of a member of the group choosing $a = 1$, $a = -1$ and $a = 0$, J_1, J_{-1} and J_0, respectively, who chooses to be transparent vis-à-vis an agent v. Vis-à-vis an agent j with type v such that $a_j = -1$, agent i's reputational payoff when $a_i = +1$ is $\Gamma^1(v, v^*)$ if his action is transparent, and, when joining a safe space,

$$\Gamma_x^1(v, -v^*) \equiv \frac{[2F(v^*) - 1]\Gamma^0(v, v^*) + x[1 - F(v^*)]\Gamma^1(v, v^*)}{[2F(v^*) - 1] + x[1 - F(v^*)]} \geq \Gamma^1(v, v^*),$$

using Assumption 4(iv)$'$ ($r(\tilde{v},v)$ is decreasing in $\tilde{v}$ for $v \le -v^* \le \tilde{v}$).

Similarly, agent i's reputational payoff vis-à-vis agent j with $a_j = 0$ is $\Gamma^1(v,v^*)$ if his action is transparent, and, when joining a safe space,

$$\Gamma^0_x(v,-v^*) \equiv \frac{[2F(v^*)-1]\Gamma^0(v,v^*) + x[1-F(v^*)]\Gamma^1(v,v^*) + xF(-v^*)\Gamma^{-1}(v,-v^*)}{[2F(v^*)-1]+2x[1-F(v^*)]}.$$

To show that Lemma A2 (demand for reputation) also holds for the reputation as a random member of a group, consider first the disclosure of $a_i = +1$ to J_0. The overall reputational gain vis-à-vis group J_0 when joining a safe space is:

$$\frac{2F(v^*)-1}{[2F(v^*)-1]+2x[1-F(v^*)]}\int_{-v^*}^{v^*}[\Gamma^0(v,v^*)-\Gamma^1(v,v^*)]dF(v),$$

while this is positive since:

$$\begin{aligned}\int_{-v^*}^{v^*}\Gamma^0(v,v^*)dF(v) &= \frac{1}{2F(v^*)-1}\int_{-v^*}^{v^*}\Big(\int_{-v^*}^{v^*} r(\tilde{v},v)dF(\tilde{v})\Big)dF(v)\\ &= \frac{1}{2F(v^*)-1}\int_{-v^*}^{v^*}\Big(\int_{-v^*}^{v^*} r(\tilde{v},v)dF(v)\Big)dF(\tilde{v})\end{aligned}$$

Using Lemma A1, $\int_{-v^*}^{v^*} r(\tilde{v},v)dF(v)$ is concave in $\tilde{v}$, and peaks at 0. Hence we have:

$$\begin{aligned}&\frac{1}{2F(v^*)-1}\int_{-v^*}^{v^*}\Big(\int_{-v^*}^{v^*} r(\tilde{v},v)dF(v)\Big)dF(\tilde{v})\\ &\ge \frac{1}{2F(v^*)-1}\Big(\int_{-v^*}^{v^*} r(v^*,v)dF(v)\Big)[2F(v^*)-1] = \int_{-v^*}^{v^*} r(v^*,v)dF(v)\end{aligned}$$

On the other hand, we know that:

$$\begin{aligned}\int_{-v^*}^{v^*}\Gamma^1(v,v^*)dF(v) &= \frac{1}{1-F(v^*)}\int_{-v^*}^{v^*}\Big(\int_{v^*}^{+\infty} r(\tilde{v},v)dF(\tilde{v})\Big)dF(v)\\ &= \frac{1}{1-F(v^*)}\int_{v^*}^{+\infty}\Big(\int_{-v^*}^{v^*} r(\tilde{v},v)dF(v)\Big)dF(\tilde{v})\\ &\le \frac{1}{1-F(v^*)}\Big(\int_{-v^*}^{v^*} r(v^*,v)dF(v)\Big)[1-F(v^*)] = \int_{-v^*}^{v^*} r(v^*,v)dF(v),\end{aligned}$$

where we invoke again Lemma A1: $\int_{-v^*}^{v^*} r(\tilde{v},v)dF(v)$ is concave in $\tilde{v}$, and peaks at 0.

Next consider the disclosure of $a_i = 1$ to J_1, or equivalently here we compute the disclosure of $a_i = -1$ to J_{-1}. We need to show:

$$\begin{aligned}&\int_{-\infty}^{-v^*}\frac{1}{1-F(v^*)}\Big(\int_{-\infty}^{-v^*} r(\tilde{v},v)dF(\tilde{v})\Big)dF(v)\\ &\ge \int_{-\infty}^{-v^*}\Gamma^1_x(v,-v^*)dF(v),\end{aligned}$$

it suffices to show that for $x = 0$

$$\int_{-\infty}^{-v^*} \frac{1}{1-F(v^*)}\Big(\int_{-\infty}^{-v^*} r(\tilde{v}, v) dF(\tilde{v})\Big) dF(v)$$
$$\geq \int_{-\infty}^{-v^*} \Gamma^0(v, v^*) dF(v),$$

which is guaranteed by Assumption 4(iv)$'$.

Finally, consider the disclosure of $a_i = 1$ to J_{-1}. We show:

$$\Gamma^1_x(v, -v^*) \geq \Gamma^1(v, v^*),$$

using the fact that $r(\tilde{v}, v)$ is decreasing in $\tilde{v}$ for $v \leq -v^* \leq \tilde{v}$.

The proof of existence of an equilibrium and it's characters follows the lines of the proof of Proposition A1. The reputational payoffs for an agent choosing $a_i = 1$ and opting for a safe space ("s") or transparency ("t") or choosing $a_i = 0$, are (the payoffs for $a_i = -1$ are obtained by symmetry):

$$\begin{cases} R^s_1(v^*, x) & \equiv & \int_{v^*}^{+\infty} \Gamma^1(v, v^*) dF(v) dF(v) + \int_{-v^*}^{v^*} \Gamma^0_x(v, -v^*) dF(v) + \int_{-\infty}^{-v^*} \Gamma^1_x(v, -v^*) dF(v) dF(v) \\ R^t_1(v^*) & = & \int_{-\infty}^{+\infty} \Gamma^1(v, v^*) dF(v) \\ R_0(v^*, x) & \equiv & \int_{v^*}^{+\infty} \Gamma^1_x(-v, -v^*) dF(v) + \int_{-v^*}^{v^*} \Gamma^0_x(v, -v^*) dF(v) + \int_{-\infty}^{-v^*} \Gamma^1_x(v, -v^*) dF(v). \end{cases}$$

Now, define:

$$\begin{aligned} S(v^*, x) &\equiv v^* - c + R^s_1(v^*, x) - R_0(v^*, x) \\ &= v^* - c + \int_{v^*}^{+\infty} \Big[\Gamma^1(v, v^*) - \Gamma^1_x(-v, -v^*)\Big] dF(v) \end{aligned}$$

denote the net benefit from acting in a safe spaces and

$$T(v^*, x) \equiv v^* - c + R^t_1(v^*) - R_0(v^*, x),$$

denote the net benefit from acting transparently. The rest of the proof is entirely similar to the proof of Proposition A1 for the existence and characterization of the symmetric equilibrium. ■

H Proof of Proposition 5

(i) An increase in right-wing polarization boosts the right-wing safe space by increasing both Θ and M^+ and making it more worthwhile to join that space. Perhaps less intuitively, the right-wing safe space, say, also expands with left-wing polarization. This is due to an heightened suspicion/amalgam effect: In particular, the right-wing safe space's out-group is perceived as more left-wing due to the increased polarization on the left (a decrease in M^-), generating more hostility for the out-group within the right-wing safe space.

(ii) and (iii) Let us begin with the mechanical (composition) effect of an increase in ρ. Keeping cutoffs constant, the right-wing safe space expands when right-wing extremism does, but not when right-wing ideas are better accepted by non-activists. The left-wing safe space is unaffected in either case.

More interesting is the impact on the image benefit of right-wing activism (when joining a safe space):

$$\frac{\partial}{\partial\rho}\Theta(v^*;\rho)\big[M^+(v^*;\rho) - M^-(v^*;\rho)\big]$$

where $\Theta(v^*;\rho) \equiv \int_{v^*}^{+\infty}\theta(v)dF(v;\rho)$ is the judgment-intensity parameter in a safe-space equilibrium and $M^+ - M^-$ is the inference benefit.

For a surge in right-wing extremism, $M^-(v^*;\rho)$ is invariant when ρ increases, while $M^+(v^*;\rho)$ and $\Theta(v^*;\rho)$[51] increase. This leads to a decrease in v^*. Similarly, the amalgam effect implies that ${}^*v$ increases as $M^+({}^*v;\rho)$ increases. Overall, a surge in right-wing extremism boosts both safe spaces.

Suppose now a surge in the acceptance of right-wing ideas. There is no composition effect. When ρ changes, $\Theta(v^*;\rho)$ and $M^+(v^*;\rho)$ are unaffected, but $M^-(v^*;\rho)$ increases. Therefore v^* increases. Similarly, $\Theta^-({}^*v;\rho) \equiv \int_{-\infty}^{{}^*v}\theta(v)dF(v;\rho)$ and $M^-({}^*v;\rho)$ are invariant, while $M^+({}^*v;\rho)$ increases, leading to an increase in ${}^*v$.

I Proof of Proposition 6 (dynamics)

(a) *Safe spaces*

Consider the following behavior: For all i and τ,

$$a_{i,\tau} = \begin{cases} +1 & \text{if } v_i \geq v^s \\ 0 & \text{if } -v^s < v_i < v^s \\ -1 & \text{if } v_i \leq -v^s \end{cases}$$

where v^s is the static cutoff. I must specify what happens when the agent deviates intertemporally from the equilibrium path. I assume that the beliefs correspond to the static beliefs corresponding to the audience's information about i's current behavior.[52]

[51] $\int_{v^*}^{+\infty}\theta(v)dF_\rho(v) = \theta F_\rho|_{v^*}^{+\infty} - \int_{v^*}^{+\infty}\theta'(v)F_\rho(v)dv$. The first term on the RHS is equal to 0 and the second is positive ($\theta' > 0, F_\rho \leq 0$).

[52] For example, if $a_{i,\tau} = +1$, date-τ members of J_1 attribute beliefs $\hat{v}_{i,\tau+1} = M^-(v^s)$ if $a_{i,\tau+1} \neq +1$ and they receive no information about i's behavior and $\hat{v}_{i,\tau+1} = M^-(-v^s)$ if $a_{i,\tau+1} = -1$ and i discloses his behavior (which won't be optimal). Similarly, if $a_{i,\tau} \neq +1 = a_{i,\tau+1}$, members of J_1 infer $\hat{v}_{i,\tau+1} = M^+(v^s)$. These beliefs can be made on-the-equilibrium-path by positing that each agent's type remains the same

With such beliefs, the static behavior is optimal in each period for all v_i. And so the static behavior is also an equilibrium of the repeated game.

(b) *Transparency*

Suppose now that h is large so that transparency prevails. Let $M(v_1, v_2) \equiv E_F[v|v_1 \leq v \leq v_2]$. We look for an equilibrium with consecutive cutoffs $\{v_K^t, \cdots, v_k^t, \cdots, v_0^t\}$ with $v_K^t > v^t$ and $v_0^t \in (c, v^t)$, converging monotonically and from above toward cutoff v_∞^t (given by $v^t - c + R^t(v^t) = R^t(0)$):

$$c < v_0^t < v_1^t < \cdots < v_k^t < \cdots < v_K^t.$$

So v_K^t is the date-0 cutoff, and v_k^t the cutoff at $\tau = K - k$. The sequence satisfies:

$$v_k^t - c + R(M(v_k^t, v_{k+1}^t)) = (1-\delta)R(0) + \delta[v_k^t - c + R(M(v_{k-1}^t, v_k^t))] \tag{A.10}$$

and

$$v_0^t - c + R(M(v_0^t, v_1^t)) = R(0), \tag{A.11}$$

with the convention that

$v_{K+1}^t = +\infty$ (so $R(M(v_K^t, v_{K+1}^t)) = R(M^+(v_K^t)))$. Condition (A.10) says that type v_k^t is indifferent between acting now and being pooled in bucket $[v_k^t, v_{k+1}^t]$ and waiting one period, earning a neutral reputation for that period but forgoing the net benefit of acting, and acting for the next period onward and being put in bucket $[v_{k-1}^t, v_k^t]$, which commands a better reputation than bucket $[v_k^t, v_{k+1}^t]$.

A continuous-time example

Suppose that type $v \in (c, +\infty)$ starts being active at time $\tau(v)$. The equilibrium is separating in types in that range and $\tau' < 0$; conversely let $v(\tau)$ with $v' < 0$ denote the type that is active from τ on. Indifference yields the following differential equation, letting i denote the rate of interest:

$$[v(\tau) - c]d\tau = \frac{R'(v(\tau))\frac{dv}{d\tau}}{i}d\tau.$$

The LHS represent the loss of waiting between τ and $\tau + d\tau$. The RHS capture the gain in reputation $R(v(\tau + d\tau)) - R(v(\tau))$, discounted until the end of the horizon. Inverting this yields

$$\frac{d\tau}{dv} = \frac{R'(v)}{i(v-c)}. \tag{A.12}$$

For example, for the *maximum norm* ($R(\hat{v}) = -\mu(V + \hat{v})$),

$$\frac{d\tau}{dv} = -\frac{\mu}{i(v-c)}.$$

from one period to the next with probability $1 - \lambda$ and is redrawn from distribution $F(.)$ with probability λ, in the limit as $\lambda \to 0$. More generally, whenever a defection from activism is perceived by the in-group as meaning $\hat{v} \leq v^s$, the deviation is not profitable from Assumption 4.

And so, given that $\tau(V) = 0$,

$$\tau(v) = \frac{\mu}{i} \log\left(\frac{V-c}{v-c}\right)$$

($\tau(c) = +\infty$).

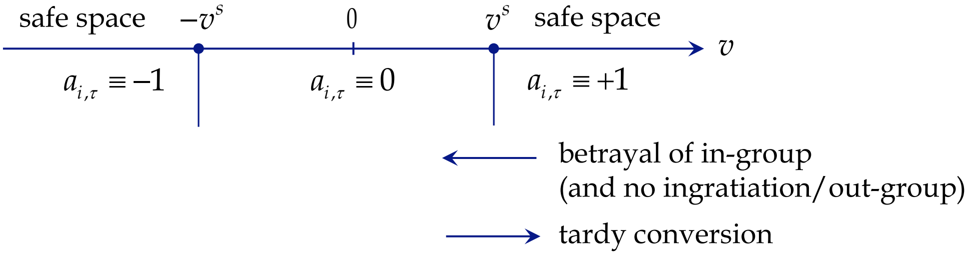

Figure 3: Low hiding cost

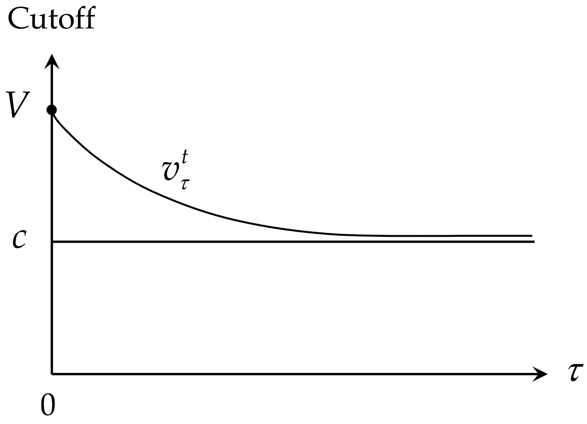

Figure 4: Dynamics under high hiding cost, the maximum norm and continuous time

J Outings and coming outs

To formalize outings, we focus on a simplified version of the model in which a fraction $(1-\alpha)$ of the population (the "moral majority") has type 0 and expresses hostility toward the fraction α of the population (the "community") who engage in an "undesirable" activity and has valuation $v > 0$ for it (we take α as fixed, unlike in the rest of the paper; this will be the case if v is sufficiently large). If known, the frowned-upon activity induces image on the members of the community

$$\begin{cases} -\mu(v+w) & \text{with probability } z \\ -\mu v & \text{with probability } 1-z. \end{cases}$$

The idea is that with probability $1 - z$, members of the community are not so different from the moral majority, while with probability z they are perceived as a different, hostile bunch ($w > 0$).

I posit that in the former case but not the latter, there exist members of the community known to the moral majority and so their outing shows that the moral majority and the community are not that different.

Let $\mu(1 - \alpha)$ denote the image concerns of an ordinary member of the community vis-à-vis the moral majority; similarly let $\mu_H(1 - \alpha)$ denote the image concerns of known members of the community.

Outing known members brings a gain equal to $\mu(1 - \alpha)\alpha z w$ to the ordinary members. By contrast, absent an outing, the known members would not have voluntarily come out if

$$-\mu_H(1 - \alpha)\alpha(v + zw) - h \geq -\mu_H(1 - \alpha)v.$$

Having imposed an outing on known members, it is an equilibrium for ordinary members to be transparent (come out) if

$$\mu(1 - \alpha)v \leq h.$$