Safe Spaces: Shelters or Tribes?*

Jean Tirole†

This draft: October 19, 2023
First draft: June 15, 2023

Abstract: The paper develops a framework for thinking about social behavior with regards to divisive issues (politics, religion, sexuality, antagonistic social views...). 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 then 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 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, Karine Van der Straeten, Glen Weyl, and participants at presentations of preliminary ideas at the AEA meetings, the IOG-BFI meeting, the NBER SI IT and Digitization workshop, Northwestern, Princeton, SITE, the Stony Brook Game Theory conference, and TSE 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

Motivation. New technologies, from AI to facial recognition, smart phones, smart glasses, GPS trackers, drones, and social networks, have increased the cost of concealing our behavior. Social norms relative to doxing, outing, and celebrity chasing have also evolved in the direction of increased disclosure.[1] Will transparency make us experience the life we aspire to? Will we look for such publicity or will we rather retreat in the comfort of a secluded environment, and if so, what kind of environment? Can regulations and platform policies address potential concerns or to the contrary exacerbate them?[2] The paper attempts to shed light on these key societal issues.

The costs and benefits of transparency hinge on individuals' reputational concerns. In this respect, lab-and-field evidence has consistently confirmed the theoretical prediction that giving a socially valued behavior more visibility makes it more prevalent.[3] This evidence however has been obtained for consensual environments, in which there is broad agreement as to what represents "good" and "bad" behavior: A vast majority of people will view selfishness, pollution or crime as bad and charitable contributions or public good provision as good.

Numerous behaviors however are appreciated by part of the audience and adversely assessed by others. Take corrida or boxing attendance, or the religious slaughtering of

[1]Collection services companies build on mobile and Wi-Fi data to allow others to know about an individual's private life. Hackers and other vigilantes "dox" participants in demonstrations and e-mail writers to expose them publicly. Rogue employees of telephone companies share data about political activities. Media publish private information about celebrities. And ordinary people's electronic word of mouth or web posting provide individual-level information regarding others' personal choices (Acemoglu et al 2022, Bergemann et al 2022 and Choi et al 2019). Even a community itself may join the disclosure game by outing a member.

Conversely, the state affects the cost of preserving privacy in many ways. For example, data regulations, from constraints on the complexity and clarity of the consent forms to rules governing the resale of data, aim at making it simpler for the citizen to protect themselves from the collection by platforms and governments of information they want to keep private. Europe for instance ruled on the right to oblivion (2014 European Court of Justice), privacy protection (2016 GDPR) and algorithms (2021 Artificial Intelligence Act). Similarly, the right of a person and the person's property to be free from unwarranted public scrutiny hinges for instance on the notion of "public space" and "editorial use". Interesting questions arise when First Amendment rights of free speech and free press clash with defamation law. Social norms also impact the attention individuals must pay to protect their privacy; until recently, it was considered bad taste for a French media to expose a politician's private life, while that of American ones has long been "fair game".

[2]An example of the creation of trusted spaces for communities around shared interests is Facebook's recommendation system. A 2016 internal Facebook presentation on extremism in Germany stated that "64% of all extremist group joins are due to our recommendation tools."

[3]See Ashraf-Bandiera (2018) and Bursztyn and Jensen (2017) for overviews of this literature. References include Freeman (1997), Ariely et al. (2009), Ashraf et al. (2012), Bursztyn et al. (2020) for charitable contributions, Algan et al. (2016) 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 (2021) 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 leads to 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. 2023).

an animal. A fraction of the population may infer good traits (attachment to cultural roots, conviviality) while others will find the behavior repulsive, a signal of bad taste or selfishness.[4] Our religious or political views, our sexual orientation, or our attitudes towards abortion, breastfeeding, surrogate motherhood, transgender rights, veganism, GMOs, vaccines, medically assisted reproduction, or social roles are divisive: frowned upon by some and approved by others. Cultural issues in countries such as the US have taken centre stage as key matters of contention.[5]

Importantly, the expansion of the public sphere is not random. The selective relationships of our private sphere are biased towards like-minded individuals; the new sharing of information may expose our behavior to a less like-minded audience, with different implications for our reputational concerns.

Modeling choices. Section 2 provides a framework for thinking about the endogeneity of our private sphere in environments in which issues are divisive. To capture behaviors for which agents do not agree on the proper way to act (the first signature of divisive issues), I assume that each agent i selects an action $a_i \in \{-1,\ 0,\ +1\}$; action $a_i = 0$ is interpreted as "not acting"/"staying neutral"/ "not picking a camp" and $a_i \in \{-1, +1\}$ as engaging in activism. One interpretation of the "not acting" choice is that it reflects an agnostic/non believer/moderate stance and plausible deniability (lack of time, other interests...).

The individual's type v_i determines the non-reputational benefit, $v_i a_i$. The non-reputational payoff is $va_i - c|a_i|$, where $c \geq 0$ is a cost of acting, so individuals with stronger convictious have higher incentives to engage. Type v_i is drawn from a symmetric and unimodal distribution $F(v_i)$ with support $(-\infty,\ +\infty)$: For any action $a_i \in \{-1,\ +1\}$, some attach a negative value to it, while others view it favorably. This environment is a good approximation for the above-mentioned realms of politics, religion, secularism, and in many countries and epochs sexual orientation, abortion, lifestyles, or wealth and income.

To formalize the idea that divisiveness affects behavior and capture the second signature of divisive issues (agents think carefully about whom they disclose their behavior to), I start from bilateral reputations and posit that agent i values her reputational payoff $r(\hat{v}_{ji}, v_j)$ with agent j. The novelty is that the reputation is in the eyes of the beholder, in two ways: First, unlike in standard models of prosocial behavior (e.g. Bénabou-Tirole 2006) or models of conformity (e.g. Bernheim 1994), agents differ in their appreciations of a fellow agent's behavior: agent i's reputation depends not only on her behavior, but also on agent j's type v_j. Second, a central aspect of the study is that the visibility of agent i's action to agent j endogenously depends on the latter's type, even though this

[4]Horizontal traits often result from multiple vertical ones that are weighted differently by different onlookers. Opponents to corridas or boxing probably value positively attachment to cultural roots and conviviality, but they put more weight on other dimensions.

[5]We 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, if the warm glow v refers to trying to inflect public policy by demonstrating, then joining a safe space would defeat the purpose.

type is private information: r depends on $\hat{v}_{ji}$, the expected type of agent i conditional on whatever agent j observes about her behavior. Agent i engages in self-presentation, i.e. manipulates the visibility of her behavior, in an audience-contingent manner.

The reputation function $r(\hat{v}_{ji}, v_j)$ is assumed to satisfy three reasonable properties besides symmetry: Ceteris paribus, the agent first wants to limit perceived taste dissonance with her 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 our two lead examples: the first involves a positional image model, in which reputation acquisition is a zero-sum game. The second takes r to depend negatively on the distance between $\hat{v}_{ji}$ and v_j, as measured by the L^p-norm. Either way, agent i's reputational payoff is taken to be the sum of bilateral reputational payoffs $r(\hat{v}_{ji}, v_j)$ with other agents j.

Under these assumptions, social welfare is maximized in the fictitious "full-privacy case", in which the individual's behavior is observed by no-one but the individual (this case never results from equilibrium behavior). This is so for two reasons. First, behavior is then authentic,[6] i.e. unencumbered by social pressure. Second, the lack of leakage about individual behavior implies that all have the same reputation. The concavity of image payoffs in the reputation implies efficiency in that dimension too.

Analysis. I first conduct a thought experiment and analyze an agent ("agent i", she)'s demand for privacy/transparency by looking at her desired self-presentation when the latter is costless for the agent. Only when acting ($a_i \in \{-1, +1\}$) does agent i have something to disclose. The demand at this stage is assumed to be fully contingent on the action selected by each agent j in the audience (but not on his type, which is not directly observable). A simple result obtains: the agent ideally discloses her behavior to other, like-minded agents choosing the same action as she does (her "in-group"), but does not disclose it to anybody else (the "out-group").[7] Put differently, the agent retreats in a safe space of like-minded individuals, and hides her behavior from others.

This simple result however raises three questions, studied in Section 3: (a) Given that the disclosure of actions by other agents is also strategic, is such a contingent disclosure by agent i feasible (the above information-design exercise must be an equilibrium information design and therefore measurable with respect to agent i's equilibrium information)? (b) What does the equilibrium look like (how authentic is the agents' behavior?) and how is it implemented in practice? (c) Are there image externalities?

Desired image and equilibrium in the absence of hiding cost. The answer to question (a) is straightforward. While agent i may not observe the exact action of agent j if the latter belongs to i's out-group, this lack of information is inconsequential as agent i's

[6] "Authenticity" refers to the agent's behavior in the absence of reputational concerns.

[7] Thus, for an agent picking $a_i = +1$, the in-group consists of other agents j selecting $a_j = +1$, while the out-group is composed of the agents selecting actions -1 or 0.

unconstrained optimal policy is to not disclose to anyone in her out-group. Furthermore, information about each other's behavior within an in-group can be shared with in-group members, and indeed the group's members desire to do so. This also provides a clue to answering (b): In the absence of self-presentation cost, the equilibrium always involves the formation of "safe spaces": Acting takes place in an environment of like-minded peers who pick the same behavior; this safe space may be physical (a political party, a religious building, a masonic lodge, a bullfight ring, a secret society...) or virtual (I develop an interpretation in terms of "repositories", which share similarities with a Facebook group). The ability to signal to a select group of sympathetic individuals implies that the individual acts more often than authenticity would command.

Regarding (c), the agent is afraid to act when her behavior is made visible to the wider audience. But safe spaces, which liberate her action, create image externalities. Moderates do not pick a camp and are viewed with suspicion by the two camps, as they are informationally pooled with the enemy camp. The consequence of this conflation (or amalgam) effect is that there is a strong incentive to take side and then enter a safe space. Behavior is not authentic, but in the opposite direction relative to the transparency case: The suspicion makes agents too prone to act.

Hiding costs. In practice, agents incur hiding or self-presentation costs to join a safe space. The first such cost relates to a reduced use of the public space. Reproved sexual minorities cannot enjoy the public space together, drug users or aborting women resort to costly and untrustworthy providers, freedom of speech may be hampered, etc. The second cost comes from the morphing of one's social graph. Creating a safe space may imply renouncing diversity in one's choice of friends and focusing on like-minded individuals one can trust not to disclose one's behavior publicly, either through empathy or because they are afraid of retaliatory outing.[8] To the extent that this means forgoing desirable relationships or that diversity is valued per se, joining a safe space is costly.

For hiding costs below some threshold, the equilibrium still involves safe spaces and agents incur wasteful hiding costs. For high hiding costs, transparency prevails. In a region in between, the equilibrium exhibits mixing between the joining of a safe space and transparency. But, as we have seen, transparency is also costly, as it makes agents afraid of undertaking an activity they like but is controversial.

When reputations are redistributive, i.e. have no aggregate image consequences (positional image), transparency is preferable to safe spaces. Reputation stealing (the conflation loss incurred by neutral agents) allowed by the possibility of retreat into a safe space creates a urge to take side and avoid being simultaneously suspected by both sides, thereby destroying authenticity of behavior; wasteful hiding costs such as a reduced use of public spaces may further be incurred when joining a safe space. The situation is differ-

[8] As Simmel (1906) notes in his discussion of the "duty of reticence", *"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."*

ent when the reputational pattern has aggregate consequences. Under the L^p norm, high levels of perceived conflict is socially costly (say, they may trigger high hostility, discrimination or verbal or physical violence); safe spaces then allow one to be authentic while protecting oneself from extreme hostility. For example, for the maximum norm (obtained as $p \to +\infty$), a safe space outcome socially dominates transparency.

Finally, Section 3 studies the impact of polarization (there are fewer moderates and more extremists) on behavior. Polarization is captured by a rotation of the distribution of types. Polarization raises the popularity of safe spaces for two reasons, the first of which is mechanical: if the equilibrium is a safe-space one, safe spaces population grows as individuals become more opiniated. More interestingly, incentives to join a safe space are altered. For instance, as right-wingers become more opinionated, the opinion of the right-wing activists matters more. Furthermore, the perceived type differential between right-wing activists and their outgroup increases. Departing from symmetric distributions, I also consider "one-sided polarization" in which only one of the sides becomes more radical. The right-wing individuals becoming more extremist 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 provides several extensions and applications. Section 4.1 studies the dynamic extension of the static model. The key insight here is that a safe-space outcome (which as we have seen obtains for low binding costs) is still an equilibrium in the repeated-action extension. The intuition is that quitting a group and stopping to take side does nothing to ingratiate the agent with the out-group (which does not observe the deviation under a safe-space configuration), while it is frowned upon by the in-group; conversely, a late convert is not viewed as favorably by the in-group as an early one. In contrast, for high hiding costs (so transparency obtains), the dynamic outcome departs from the static one and involves "Coasian features"; the agent can build a reputation for moderation by remaining passive during a few periods. The audience then "knows" that she is not an extremist and over time becomes more and more tolerant of her acting; that is, the stigma from entry into activism is time-decreasing. The flip side of the coin is that neutral types are under more and more pressure to take side over time.

The very demand for privacy implies that one of the worst fears of a member of a community is to be outed. In practice, outings tend to be more frequent for high-image-concerns members (politicians, celebrities, local notables...). Section 4.2 extends the model to account for outing and for coming outs, the former facilitating the latter. The idea, consistent with some empirical evidence, is that the outing of a celebrity moves her group's outside perception toward the mainstream and reduces the hostility of the outgroup. This benefits group members, all the more that there can be fortuitous (involuntary) revelations of membership or else a desire for transparency (voluntary coming outs).

Section 4.3 analyses the case in which preserving privacy requires focusing on a social graph of like-minded peers whom one can trust. Creating a safe space then implies a

personal cost in terms of diversity (friends are selected in a smaller group and are less diversified). It also involves a one-shot cost of making new friends and abandoning old ones. These two costs arguably can be well captured by the L^1 distance between the prior distribution of types and the more restricted distribution associated with the demand for privacy. Two interesting results then emerge. First, endogenous social graphs introduce a form of strategic complementarity: if a broader set of agents retreat in safe spaces, there is more diversity in safe spaces, and the diversity and switching costs of joining a safe space are reduced, making safe spaces privately more attractive. Second, due to the switching cost, social graphs exhibit a form of hysteresis: while it is costly for agents to morph their social graph to create a safe space, safe spaces are hard to undo once the social graphs have been ghettoized.

Section 4.4 allows agent i's reputational payoff to depend on j's entire conditional distribution about i's type and not just the conditional mean (of course the two coincide if the reputational payoff is linear in i's perceived type). In that case, image is constant-sum and the analysis is the same as for the positional case mentioned earlier. Transparency is then socially optimal.

Section 5 considers additional signaling within the safe space. This extra signaling may be voluntary (and yet inefficient) as the group then descends in one-upmanship. Agents want to show they are "the true believers". As is well-known from Facebook groups and other fora inhabited by like-minded individuals, one-sided information and narratives will circulate within groups when they would not circulate if the audience were not restricted to like-minded individuals. Such behaviors add to the potential social costs that exist even in the absence of such additional channels for signaling. Alternatively, the in-group may demand some actions that would not voluntarily be chosen by members but serve the community as a whole. Whether spontaneous or coerced, such one-upmanship imposes costs either on the ingroup (e.g., costly rituals) or on the outgroup (aggressions, biased beliefs). Excluding the more radical elements (as defined by their behavior) may help the group protect itself against such costs by depriving the radicals from access to a sizeable community.

Public policies make it easier or harder to protect one's privacy. This analysis suggests that their optimal design hinges on a couple of considerations. Protecting the agents' ability to join a protective safe space is desirable a) for behaviors that are sensitive and generate social-value-destroying behaviors (hostility, insults, discrimination, beatings, pogroms...) and b) if the rule of law and social norms are weak and do a poor job at protecting the population against such behaviors. Pushing toward transparency are c) reputation externalities that are mostly redistributive/do not destroy social value, and especially d) the broader social concerns associated with both the ghettoisation of thinking and the extra posturing against the out-group that occurs in safe spaces, either naturally (one-upmanship) or enforced by the threat of exclusion or outing. Section 6 concludes with alleys for future research.

Related literature. A large literature analyses how image concerns impact the behavior of

individuals. This includes for instance Bernheim (1994) and Bénabou-Tirole (2006). As we discussed, both the conformity and the prosociality models are models of consensual behaviors (all agree on the ranking of types). So the demand for reputation under costless self presentation is independent of the audience.[9] 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.

A sizeable literature has analyzed the implications of a preference for conformity (e.g. Bernheim 1994, Manski-Mayshar 2003, Kuran-Sandholm 2008, Michaeli-Spiro 2015, 2017, Braghieri 2021). 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' actions, as in Michaeli-Spiro 2017). 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 Bénabou-Tirole). This paper 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.

My analysis of divisive issues is related to the literatures on countervailing incentives[10] and on the impact of transparency on the behavior of agents signaling to multiple audiences. In Gertner et al (1988), an informed firm signals to two uninformed audiences, the capital market and the product market. Transparency of communication may induce conflicts: for instance, the firm wants the capital market to believe that demand is high and a potential entrant that the demand is low. In Spiegel-Spulber (1997), a firm wishes to signal high value to capital markets to boost its market value while also signaling high cost to regulators to induce rate increases. In Austen-Smith-Fryer (2005), a high signal (education say) generates higher wages but also leads to more group rejection. In Bursztyn et al (2017), single women refrain from volunteering for leadership roles or asking for a promotion that might help their career, by fear of sending a negative signal to the marriage market. In these papers, the equilibrium pattern of signaling differs from that in this paper, due to the difference in signaling space; also I allow the degree of transparency to vary endogenously.

Bar-Isaac and Deb (2014) study the impact of transparency in the context of an infinitely repeated, two-audience setting with, as is the case here, three choices (left, middle, right). The agent has two possible types (L and R) and selects two actions per period. If each audience observes only one action, this action may cater to the relevant

[9] Recent papers have extended our knowledge on such signaling incentives. For example, the sender can garble the performance signal (Ball 2022). 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).

[10] Lewis-Sappington (1989), Maggi-Rodriguez-Clare (1995).

audience (left or right); if each audience observes both actions, the compromise action is more likely. Payoffs are action-based, not belief-based (still, under separate observations, beliefs help infer the unobserved action). They show that reputation concerns increase the sender's welfare under transparency and reduce it under separate observations by the two audiences. Bouvard and Levy (2017) consider a certifier who must attract both sides of the market (see also Frenkel 2015). Buyers demand a high accuracy, but too high an accuracy may repel weaker sellers. They look at the building of reputation concerning certification accuracy. Besides the differences in context and modeling, a major focus of the paper is the endogenization of disclosures, indicating when there is a demand for privacy and leading to policy recommendations.

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. Philosophers and economists share the idea that people distort their public actions due to social-reputational payoffs.[11] "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.[12] The paper derives insights about how the endogeneity of the public and private spheres in our lives affects our well-being

[11]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.

[12]Erich Fromm (1941) seemed to be aware of the tension between the rosy view of authenticity and such socially detrimental behaviors. He did not see authenticity as a mere rejection of the expectations of others and allowed authenticity to reflect those cultural norms that appear appropriate (see later for an analysis of when the strengthening of a norm improves or reduces welfare, which may provide some foundations for their "appropriateness"). See, e.g., the Wikipedia entry on "Authenticity (philosophy)" for an account of the broader debate among philosophers concerning the notion of authenticity. The tension between the positive (emancipation) connotation of "authenticity" in the private sphere and the public discourse on transparency as a factor of social harmony can be resolved by considering the magnitude of externalities. The positive connotation is vindicated whenever the activity involves - or is perceived by proponents of authenticity to involve- a low externality: norms of etiquette, excessive attention to self-presentation, obedience to majoritarian views on religion, sexuality or soft drugs, conformity with stereotypes and social roles, etc. Larger externalities by contrast call for social accountability, which absent extrinsic incentives is brought about by transparency.

in a divisive-issue context.

2 Divisive behaviors

2.1 Modeling

The set of agents has mass 1 and agents are indexed by i, j. Each agent i picks an action $a_i \in \{-1, 0, +1\}$. We assume that people disagree as to what is "moral" or "immoral", "good" or "bad", "right" or "wrong". We here have in mind political, religious and broader societal issues on which there is no consensus. Agent i can stay passive/neutral ($a_i = 0$) or act/pick a camp ($a_i = -1$ 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 it, etc.). Agent i has privately-known type (value, ideology) v_i and non-image payoff from her action a_i:

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

Preference heterogeneity. The basic model posits a common knowledge[13] 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. Symmetry around the mode 0 in part captures the lack of consensus. We assume that the distribution has a mean (necessarily 0 given symmetry).[14]

Image concerns. We assume that evaluations are in the eyes of the beholder: Let $r(\hat{v}, v)$ denote the (pairwise) reputational payoff of an agent with another agent who has type v and attributes expected value $\hat{v}$ to the former agent; that is, if agent j has type v_j and, given her information, estimate $\hat{v}_{ji}$ of agent i's type, agent i has reputational payoff with agent j equal to $r(\hat{v}_{ji}, v_j)$.[15] We assume that r is twice continuously differentiable and

[13]For simplicity, we take the perceived sense of polarization to be the true one, and rule out pluralistic ignorance. Bordalo et al (2022) on the basis of US data argue that individuals overestimate the actual polarization in the population. The possibility of pluralistic ignorance can be captured in the asymmetric-distribution version of my model for example by introducing imperfect information about the radicality of opinions on the other side. This would open the possibility of correcting beliefs through norm-based interventions or familiarity with the other side (as in Allport's 1954 contact hypothesis or Levy (2021)'s observation that exposure to counter-attitudinal news decreases negative attitudes toward the opposing side). Similarly, we take individual preferences as given, when more generally they may be shaped by one's social network (Algan et al 2019, Golub-Jackson 2012).

[14]As is the case for distributions with a bounded support, or standard distributions with unbounded support such as the normal distribution, or the symmetrized exponential or Pareto distributions. The existence of a mean requires that the distribution's tails not be too thick. For example, the Cauchy distribution ($f(v) = 1/\pi(1 + v^2)$ in its symmetric form) does not have a mean. The existence of a mean implies that $\lim_{v^* \to +\infty} \int_{v^*}^{+\infty} v dF(v) = 0$.

[15]We thus anchor agent i's reputational payoff vis-à-vis agent j to her *expected* type given j's information about her behavior. In that sense, agent i is viewed as representative of a perceived group. An alternative hypothesis, entertained in Section 4.4, 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),

denote $r_1 \equiv \partial r/\partial \hat{v}$, $r_{11} \equiv \partial^2 r/\partial \hat{v}^2$, etc.

The overall reputational payoff of agent i is in a first step assumed additive:

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

Suppose that agent i's audience can be divided into groups J with the same information about, and therefore assessment of agent i's type within each group J (that is, $\hat{v}_{ji} \equiv \hat{v}_{Ji}$ is the same for all $j \in J$); if group J has mass m_J and conditional distribution $F_J(v)$ (satisfying $\Sigma_J m_J F_J(v) \equiv F(v)$ for all v), then R_i can be rewritten as

$$R_i = \Sigma_J m_J \int_{-\infty}^{+\infty} r(\hat{v}_{Ji}, v) dF_J(v).$$

Payoff functions and equilibria. Agent i may incur a "self-presentation cost" h_i (a prominent application will refer to "hiding"). Self-presentation, technically an exercise in audience-contingent (non) disclosure, will be application-specific and discussed later on. We assume, though, that h_i is independent of agent i's type (but it is contingent on the choice of her action and the selected disclosure strategy). Summing up, agent i's utility is

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

All along, agent i's type, v_i, will not be directly observable by other agents. At most, the latter will observe a signal about her action a_i. This implies that R_i will, like h_i, be action a_i-, but not type v_i-dependent.

Agent i, when choosing action $|a_i| = 1$, can affect the visibility of her behavior to others, with an eye on maximizing $R_i - h_i$. In contrast, passive agents ($a_i = 0$) face no such disclosure decision as they have nothing to disclose, and accordingly never incur a hiding/non-disclosure cost. The separability of u_i allows us to make the Markov assumption that all types choosing a given action a_i, with $|a_i| = 1$, make the same choice of self-presentation.[16]

We will focus on *symmetric* equilibria (Proposition 4 below shows that under weak assumptions, there is no asymmetric equilibrium, and a unique symmetric one). The sorting condition implicit in the definition of u_i then implies that there exists a cutoff $v^* \geq 0$ such that

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

An agent's in-group (out-group) is endogenously defined as the set of agents who make the same (a different) choice of a.

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.

[16]The optimality condition only implies that the average $r(\hat{v}_{ji}, v_j)$ over the population of agents j depends on a_i, but not on v_i.

2.2 Assumptions on image concerns

The reputational payoff is assumed to satisfy:

Assumption 1 *(symmetry). For all* $(\hat{v}, v)$,

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

Assumption 2 *(distaste for dissonance). Ceteris paribus, agents want to ingratiate themselves with others. Suppose that* $v > 0$.[17] *Then for all* $\hat{v} < v$[18]

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

Assumption 3 *(concavity). Perceived ideological disapproval has an increasing marginal cost: for all* $(\hat{v}, v)$,

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

Assumption 4 *(benefit from being perceived by the in-group as representative of the in-group rather than as the average type in the population). Let* $M^+(v^*) \equiv E[v|v \geq v^*]$. *An agent picking* $|a_i| = 1$ *gains from being perceived by her in-group as the mean type of the group rather than as the average type in the population: for all* $v^* \geq 0$,

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

2.3 Examples

I provide interesting illustrations of image concerns that satisfy Assumptions 1 through 4. These illustrations are selected not only for their tractability, but also for their different nature. The intensity of image concerns in these illustrations is scaled up or down by a parameter $\mu > 0$.

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 ($\theta(-v) = -\theta(v)$) satisfying $\theta(0) = 0$. Note that Assumption 3 is satisfied with equality ($r_{11} = 0$). Indeed, $r_{11} = 0$ is characteristic of the positional image paradigm.[19]

[17] By symmetry, Assumption 2 implies that for $v < 0$ and $\hat{v} > v$, then $r_1(\hat{v}, v) < 0$.
To see this, use Assumption 1. 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 2 then implies that $r_1(-\hat{v}, -v) > 0 \Leftrightarrow r_1(\hat{v}, v) < 0$.

[18] Note that we 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 $\hat{v} > v > 0$.

[19] 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 1 (applied to $\hat{v} = 0$ and $\hat{v} \neq 0$) implies that γ is symmetric and θ antisymmetric. Image is therefore constant-sum.

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 exclusive the agent's in-group (i.e., the higher v^* 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 more exclusive the in-group. And conversely for action $a_i = -1$.[20]

Illustration 2: Placating image concerns. Suppose that agents want to be perceived as close in values as possible to their audience.[21] Let such concerns be measured by the (modified) L^p-norm distance[22] between perceived type and onlooker's type, and thus satisfy

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

where $\mu > 0$ and $p \geq 1$ (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 in some applications, the agent may care more about hostile than about favorable opinions, say because hostile opinions may trigger hate and violence. The parameter μ captures the intensity of image concerns. These image concerns trivially satisfy Assumptions 1, 2 and 3 for all p. More interestingly, they also satisfy Assumption 4 for all p, as is shown in the Appendix.[23]

Lemma 1 *The positional-image reputation and (for all integers $p \geq 1$) the modified L^p-norm one satisfy Assumptions 1-4.*

Alternatively, we can consider non-additive forms:

[20] The antisymmetry of $\theta(\cdot)$ facilitates the treatment. However, it is stronger than needed. For example, the negative-v group might frown upon the authentic behavior of the positive-v one, but the converse may not hold: The positive-v group may not reject the authentic behavior of the negative-v one or feel any urge for proselytism.

[21] 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.

[22] This is an F-norm. It is homogeneous of degree p.

[23] This is straightforward when $p = 1$ or $p = 2$. For the Euclidean distance ($p = 2$),

$$\int_{v^*}^{+\infty} [r(M^+(v^*), v) - r(v^*, v)]dF(v) = \mu[1 - F(v^*)][M^+(v^*) - v^*]^2 > 0.$$

For the L^1-norm, this expression is equal to

$$\int_{v^*}^{M^+(v^*)} [M^+(v^*) + v^* - 2v]dF(v) + [1 - F(M^+(v^*))][M^+(v^*) - v^*]$$
$$\geq \int_{v^*}^{M^+(v^*)} [M^+(v^*) - 2v]dF(v) + [1 - F(M^+(v^*))]M^+(v^*) > 0.$$

The second term is positive and so is the first term because the density f is decreasing in the positive domain.

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

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

As we will show, the characterizations in this paper also hold for the true L^p norm.

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 $-\mu\big(\int_{-V}^{V} |\hat{v}_{ji} - v_j|^p dF(v_j)\big)^{1/p}$, 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 her audience. The maximum norm captures an extreme form of conflict aversion.

2.4 Demand for reputation

To grasp what Assumptions 1-4 imply for the image that the agent wants to project, we consider two polar thought experiments. In the first, the agent can do nothing to conceal her behavior from anyone. In the second, she can costlessly select whom to disclose her behavior to, subject to the measurability condition that the disclosure rule depends on the receiver's action (the receiver's type is never observable by anyone but the receiver). Both will make use of the following simple lemma:

Lemma 2 *Under Assumptions 1 and 3, for any $b > 0$, $\int_{-b}^{+b} r(\hat{v}, v) dF(v)$, which is concave in $\hat{v}$ (strictly so if $r_{11} < 0$), peaks at $\hat{v} = 0$.*

Proof of Lemma 2. We need to show that $\int_{-b}^{+b} r(\hat{v}, v) dF(v)$, which is concave from Assumption 3, peaks at 0. Its derivative at 0 is

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

Assumption 1 implies that $r_1(\hat{v}, v) = -r_1(-\hat{v}, -v)$ and so $r_1(0, -v) = -r_1(0, v)$, implying that $\int_{-b}^{+b} r_1(0, v) dF(v) = 0$. ■

A direct consequence of Lemma 2 is:

Proposition 1 *(benefit from appearing as moderate under transparency). Let $R^t(\hat{v}) \equiv \int_{-\infty}^{+\infty} r(\hat{v}, v) dF(v)$ denote the reputational benefit of an agent with homogeneous reputation $\hat{v}$ (as is the case under transparency). R^t is symmetric and concave in reputation $\hat{v}$ and peaks at $\hat{v} = 0$. It is strictly concave whenever $r_{11} < 0$ (by contrast R^t is flat at 0 in the positional image case).*

Definition 1 *(canonical equilibrium). A canonical equilibrium is such that (a) agents with $v \geq v^*$ and only they choose $a_i = 1$ (and symmetrically for agents with type $v \leq -v^*$) for some $v^* \geq 0$; (b) active agents disclose their behavior to their peers (those who choose the same action, the in-group) with probability 1 and disclose it to non-peers (the out-group, which includes passive agents) with probability $1 - x$ (so x is the probability of hiding one's behavior from non-peers).*

In a canonical 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:[24]

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

Consider a canonical equilibrium and conduct the thought-experiment in which active agent i selects whom to disclose her action $|a_i| = 1$ to (the agent must tell the truth, but not necessarily the whole truth: She cannot prove that she did not act, i.e. that $a_i = 0$). This disclosure must be measurable with respect to the action selected by the receiver, as types themselves are never observable. The thought-experiment corresponds to the case in which disclosure strategies have no impact on payoffs ($h_i \equiv 0$).

Proposition 2 *(demand for reputation). Consider a canonical equilibrium. Under Assumptions 1 through 4, and ignoring any cost of self-presentation, an agent i who selects $|a_i| = 1$ strictly prefers to disclose her behavior to her peers, and prefers not to disclose her behavior to non-peers (strictly so unless $v^* = 0$ and $x = 0$); and so $x_i = 1$.*

Proof: 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 $a_i = 1$, agent i's image with them 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

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

[24] 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 strictly prefers them to $a_i = 0$). We will then naturally assume that posterior beliefs put all weight on $v = 0$: $F(v; 0) = 0$ for $v < 0$, $= 1$ for $v > 0$.

Note that $\hat{v} \leq 0 \leq v^*$. So, from Assumption 2 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 Assumption 4.

Next, consider the disclosure of $a_i = 1$ to the passive group ($v \in [-v^*, v^*]$). Not disclosing yields reputation $\hat{v} = 0$ while disclosing leads to image $\hat{v} = M^+(v^*) > v^*$. Lemma 2 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

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

and so the gain from non-disclosure is:

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

from Assumption 2. ■

To sum up, an active agent ceteris paribus wants to share the nature of her behavior with her peers, but not with her non-peers. This will indeed be the case (and so $x = 1$) if manipulation (non-disclosure) is costless. This observation however raises two questions. First, how could such an equilibrium arise? Second, and considering the more interesting case in which manipulation is costly, how is the symmetric equilibrium determined?

3 The emergence of safe spaces and their implications

We first derive the results for the additive case (e.g. examples 1 and 2) and later generalize them to non-additive reputational payoff functions.

3.1 Costless self-presentation

Proposition 2 implies that if there is no cost of self-presentation ($h_i \equiv 0$), agents prefer to disclose only to their in-group in a canonical equilibrium ($x = 1$). Note that agents in $J_1 \equiv \{j | a_j = 1\}$ then cannot tell apart agents in J_0 from those in J_{-1}, as the latter do not share their behavior with them; and so one might be concerned about the measurability of their desired disclosure strategy as analyzed in Section 2.4. However, this strategy specifies

the same strategy –no disclosure– for both groups J_0 and J_{-1}. And so measurability is not an issue. We now define the notion of a safe space.[25]

Definition 2 *(safe space). A safe space is an environment in which an agent can act* ($a_i = 1$ *or* $a_i = -1$) *and be observed by those who choose the same action (her "in-group" or "peers"), but by no-one else.*

In the following, silo or safe space information ("s") will refer to observability by the in-group audience only, while transparency ("t") will correspond to observability by all. We provide two interpretations of a safe space:

Activity-based visibility. A natural interpretation of a safe space is "activity-based visibility". Indeed, relationships in our private sphere are biased toward like-minded individuals. This bias arises whenever taking part in an activity also defines who observes that one takes part in the activity: a safe space such as a family house, a church, a corrida, a club, an ethnic neighborhood or a political party is attended only by peers.

Virtual safe spaces. A more abstract, alternative foundation for activity-contingent visibility is the introduction of "repositories", to which agents can credibly disclose in a verifiable way their choice when they act. A repository is a two-sided platform defined by (a) the information that is trusted to the repository through voluntary disclosure by agents; and (b) it access policy, defining who is entitled to obtain the repository's information. There is free entry into the design of repositories. From Proposition 2, an agent choosing $|a_i| = 1$ wants to disclose to the in-group, but not to the out-group. Under free entry into the repository industry, repositories will cater to this desire and therefore create safe spaces of like-minded individuals. While the concept of repository is abstract, an analogy with Facebook groups may be useful. Facebook recommends to its individual users groups of users who are like-minded. Facebook thereby contributes to what are de facto safe spaces.

The joining of a safe space when acting will be an equilibrium outcome under either activity-contingent visibility or when agents can share their information within repositories, whenever joining is costless.

Characterization of equilibrium in the absence of self-presentation cost

Consider first an interior equilibrium: $v^* > 0$. Because it is optimal for agent i with action $a_i = 1$ to disclose only to the in-group and the choice between $a_i = 0$ and $a_i = 1$ does not affect the information (none) held by agents in J_0 and J_{-1} about agent i, an equilibrium cutoff $v^* = v^s$ (again, "s" stands for "silo" or "safe space") is given by

$$v^s - c + \int_{v^s}^{+\infty} [r(M^+(v^s), v) - r(M^-(v^s), v)]dF(v) = 0 \tag{2}$$

[25] In practice, a safe space need not be completely safe. 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).

[We will study uniqueness later, in the more general self-presentation-cost framework.] Note that $v^s < c$ from Assumption 4.

If

$$c \leq \int_0^{\infty} [r(M^+(0), v) - r(-M^+(0), v)]dF(v), \tag{3}$$

there is a corner solution with $v^s = 0$. Social pressure then implies that all agents take sides: $J_0 = \varnothing$.

It is interesting to compare the safe-space equilibrium with two polar benchmarks:

(a) Transparency benchmark. Let us first compare this outcome with the one that prevails when agents cannot retreat in safe spaces so all agents have access to each other's behavior (or face a very high cost h of doing so). Let v^t ("t" stands for "transparency") denote the corresponding cutoff:

$$v^t - c + \int_{-\infty}^{+\infty} [r(M^+(v^t), v) - r(0, v)]dF(v) = 0 \tag{4}$$

Note that $v^t \geq c$ from Lemma 2 (applied to $b = +\infty$ and $\hat{v} = M^+(v^t)$).

(b) Full privacy benchmark. Another 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.[26] 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. Another angle at the safe-space equilibrium behavior is provided by looking at the welfare of passive agents, with types in J_0. Their payoff is equal to

$$u_0 \equiv 2\int_{v^s}^{+\infty} r(M^-(v^s), v)dF(v) + \int_{-v^s}^{v^s} r(0, v)dF(v).$$

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.[27]

Proposition 3 *(costless self-presentation).*

[26] While we 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.

[27] Assumption 1 (symmetry) implies that

$$\int_{-\infty}^{+\infty} r(0, v)dF(v) = 2\int_0^{+\infty} r(0, v)dF(v) = u_0 + 2\int_{v^s}^{+\infty} [r(0, v) - r(M^-(v^s), v)]dF(v) > u_0$$

from $M^-(v^s) < 0$ and Assumption 2.

(i) There exists a symmetric equilibrium characterized by cutoffs $\{-v^s, v^s\}$ and the emergence of safe spaces. The cutoff, if interior ((3) is not satisfied), is given by (2) and satisfies $v^s < c$. When (3) is satisfied, the equilibrium involves only two safe spaces ($v^s = 0$) and all agents taking sides.

ii) The passive agents ($v \in J_0$) suffer a negative externality in a safe-space equilibrium as they are viewed suspiciously by both sides.

(iii) By contrast, in the transparency benchmark, passive agents do not suffer such an externality; the cutoff satisfies $v^t \geq c$, with strict inequality if $r_{11} < 0$. Both the safe-space equilibrium and (if $r_{11} < 0$) the transparency benchmark depart from authentic behavior ($v^* = c$), defined as the behavior that would prevail under full privacy.

3.2 Costly self-presentation

Behaviors are often transparent when practiced in the public space. This implies that preserving one's privacy requires reducing one's use of the public space and thus involves hiding costs.[28] Reproved sexual minorities cannot enjoy the public space together, drug users or aborting women resort to costly and untrustworthy providers, freedom of speech may be hampered, etc. We capture such hiding costs in the following way: To ensure visibility solely within the community choosing the same action $|a_i| = 1$ rather than letting everyone access the information, an agent must spend hiding cost $h \geq 0$. So $h_i \in \{0, h\}$.

Consider an equilibrium in which hiding by agents choosing $|a_i| = 1$ occurs with probability $x \in [0, 1]$. This probability has the standard mixed-strategy interpretation: When hiding and being transparent yield the same image payoff after acting, a fraction x of those who have acted go to a safe space and the remaining fraction $1 - x$ does not bother. Then

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

are the expected means (from the point of view of an active agent) when not knowing anything about an agent's behavior except for the fact that she does not belong to one's in-group.

And let the reputational payoffs for an agent choosing $a_i = 1$ and opting for a safe space ("s") or transparency ("t") or choosing $a_i = 0$, be (the payoffs for $a_i = -1$ are

[28] For some activities, the hiding cost may be nil. For instance, entertaining subversive thoughts may be kept private at little cost; by contrast, sharing them within a safe space requires screening and confining relationships within a selected group of individuals perceived as reliable.

obtained by symmetry):

$$\begin{cases} R_1^s(v^*,x) & \equiv & \int_{v^*}^{+\infty} r(M^+(v^*),v)dF(v) + \int_{-v^*}^{v^*} r(0,v)dF(v) + \int_{-\infty}^{-v^*} r(M_x^+(-v^*),v)dF(v) \\ R_1^t(v^*) & = & \int_{-\infty}^{+\infty} r(M^+(v^*),v)dF(v) \\ R_0(v^*,x) & \equiv & \int_{v^*}^{+\infty} r(M_x^-(v^*),v)dF(v) + \int_{-v^*}^{v^*} r(0,v)dF(v) + \int_{-\infty}^{-v^*} r(M_x^+(-v^*),v)dF(v). \end{cases}$$

Let us obtain some useful properties under Assumptions 1-4. Note that for all $\{v^* \neq 0,\ x \neq 0\}$,[29]

$$R_1^s(v^*,x) > \max\{R_1^t(v^*), R_0(v^*,x)\}, \tag{5}$$

where $R_1^s > R_1^t$ results from Proposition 2 and $R_1^s > R_0$ from Assumption 4 (since $R_1^s(v^*,x) - R_0(v^*,x) \equiv \int_{v^*}^{+\infty}[r(M^+(v^*),v) - r(M_x^-(v^*),v)dF(v)] > 0$). Similarly, for all $v^* \geq 0$,

$$R_0(v^*,0) \geq R_1^t(v^*)$$

with strict inequality if $r_{11} < 0$.[30] Note also that, using the facts that $\partial M_x^-/\partial x < 0$ and that $r_1 > 0$ for $v > \max\{0,\hat{v}\}$,

$$\frac{\partial}{\partial x}(R_1^s(v^*,x) - R_0(v^*,x)) > 0.$$

In words, an increase in the use of safe spaces (in x) creates more suspicion on passive agents relatively to hiding active agents and raises the incentive to select $|a_i| = 1$ and not to disclose: There are *strategic complementarity in acting*.[31]

[29] For $v^* = 0$, $x > 0$, $R_1^t(0,x) = R_1^t(0)$.

[30] Assumption 1 implies that

$$\begin{aligned} R_0(v^*,0) - R_1^t(v^*) &= \int_{-\infty}^{+\infty}[r(0,v) - r(M^+(v^*),v)]dF(v) \\ &= \int_0^{+\infty}[2r(0,v) - r(M^+(v^*),v) - r(-M^+(v^*),v)]dF(v). \end{aligned}$$

Assumption 3 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 $\hat{v}$.

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

$$\frac{\partial R_0}{\partial v^*} = 2f(v^*)[r(0,v^*) - r(M_x^-(v^*),v^*)] + 2\left[\int_{v^*}^{+\infty} r_1(M_x^-(v^*),v)dF(v)\right]\frac{dM_x^-(v^*)}{dv^*} > 0$$

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 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.

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

$$\frac{\partial}{\partial x}(R_1^t(v^*) - R_0(v^*, x)) > 0.$$

Existence of an equilibrium

Let

$$S(v^*, x) \equiv v^* - c + R_1^s(v^*, x) - R_0(v^*, x)$$

denote the cutoff's net benefit from acting in a safe spaces rather than being passive and

$$T(v^*, x) \equiv v^* - c + R_1^t(v^*) - R_0(v^*, x)$$

denote the net benefit from acting transparently rather than being passive. For conciseness, we avoid equilibrium multiplicity by making:

Assumption 5 *(monotonicity). $S(v^*, x)$ and $T(v^*, x)$ are strictly increasing in v^* for all x.*

In examples 1 and 2, Assumption 5 holds if image concerns, indexed by μ, are "not too high" and (for the modified L^p norm) the distribution F has finite support. In contrast, it cannot be guaranteed under the modified L^p norm for $p > 1$ when the support of F is unbounded. Adding homogeneity, though (illustrations 3 and 4), will later allow us to show that Assumption 5 holds (for image concerns that are not too high) for the true L^p norm if F does not have fat tails, in that $f(v)v^p$ is bounded above.

Safe-space equilibrium. A safe-space equilibrium ($x = 1$) satisfies, for an interior cutoff,

$$v^* - c + R_1^s(v^*, 1) - h = R_0(v^*, 1) \tag{6}$$

and

$$R_1^s(v^*, 1) - h \geq R_1^t(v^*, 1). \tag{7}$$

From (5), a safe-space equilibrium satisfies $v^* < c$ as long as $R_1^t(v^*) > R_0(v^*, 1)$. In particular, $v^* < c$ for h small enough. The safe-space equilibrium cutoff $v^*(h)$ is strictly increasing in h from condition (6) and Assumption 5.

Transparency equilibrium. A transparency equilibrium ($x = 0$) cutoff satisfies

$$v^* - c + R_1^t(v^*) = R_0(v^*, 0) \tag{8}$$

and

$$R_1^t(v^*) \geq R_1^s(v^*, 0) - h. \tag{9}$$

Mixed-strategy self-presentation equilibrium. Such an "mixed equilibrium" ($0 < x < 1$) must satisfy:

$$v^* - c + R_1^s(v^*, x) - R_0(v^*, x) - h = 0 \tag{10}$$

and

$$R_1^s(v^*, x) - h = R_1^t(v^*). \tag{11}$$

Proposition 4 *(existence and characterization). Make Assumptions 1 through 5. 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 transparency 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 transparency equilibrium ($v^t \geq c$, with a strict inequality if the reputational payoff is strictly concave in reputation: $r_{11} < 0$).

Proposition 4 turns the standard result for consensual behaviors that a higher visibility increases compliance on its head: When behaviors are divisive, privacy encourages activism. Recall that 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 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 privacy.

Finally, the cutoff v^s under a safe space is strictly positive if $c > R_1^s(0,1) - R_0(0,1)$. This condition, which also holds if image concerns are not too strong, guarantees that there will be passive agents even for $h = 0$; if it fails, then $v^s = 0$ over an interval $h \in [0, h_0]$.

Remark (how free free speech is). Proposition 4 suggests some conjectures as to a dual impact of technology on our behavior regarding divisive issues. Technology exposes aspects of our life to a wider audience, making our behavior more cautious. Yet technology may also have freed (some form of) speech by enabling the creation of new spaces of like-minded individuals, most notably within social networks; this has restored incentives to engage (with dire consequences though, as we discuss in Section 5). This dual evolution may be observed in politics. A respectful exchange of political opinions may have become less frequent in the public space, although it remains vigorous in more homogeneous private spaces (family, friends, and, decreasingly, academia).

3.3 Welfare

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 $\mathcal{R} \equiv \int_{-\infty}^{+\infty} [\int_{-\infty}^{+\infty} r(\hat{v}_{ji}, v_j) dF(v_j)]$ $dF(v_i)$. 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 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 and the rent of more committed types:[32]

$$W(v^*, x) \quad = \quad R_0(v^*, x) + 2 \int_{v^*}^{+\infty} (v - v^*) dF(v).$$

The proof of the following proposition can be found in the Appendix.[33]

Proposition 5 *(total image payoff and welfare). Under Assumptions 1 through 4, and keeping non-image-management behavior (i.e. v^*) constant:*

(i) More information reduces total image payoff:

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

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 $W^t = W^{fp}$).

3.4 Illustrations: Positional and maximum norm image concerns

We now look at the cases of positional and maximum norm image concerns (which are somehow polar cases). The Appendix performs the computations for the Euclidean norm.

[32] We ignore benefits for the audience of having more information; such benefits would tilt the balance in favour of transparency. One difficulty lies in the social weight to be put on the audience's value of information. Improving an onlooker's information may benefit him, but hurt another onlooker with opposite views. Furthermore, if this knowledge enables the onlooker to hate or discriminate against the agent, it is unclear how much social weight one should put on the corresponding "benefit".

[33] The ranking among full privacy, full transparency, and the three equilibrium configurations in Proposition 5 extends to asymmetric distribution functions F. There are two important differences when the distribution is asymmetric. First, there are in general two asymmetric cutoffs $^*v$ and v^*. Second, there may exist hybrid equilibria, in which say the majority acts transparently while the minority hides in a safe space (Proposition 5 only compares information structures in the three types of symmetric equilibria and in the two benchmark cases). The proof in the asymmetric case mimics that for a symmetric distribution: Compute the total reputational payoff of others vis-à-vis an arbitrary audience type v. The five information structures are ranked according to the mean-preserving-spread criterion. The concavity of r in its first argument then yields the comparison: see the Appendix.

3.4.1 Positional image concerns

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,[34]

$$\begin{cases} R_1^s(v^*, x) &= \Theta(v^*)[M^+(v^*) - M_x^+(-v^*)] \\ R_0(v^*, x) &= -2\Theta(v^*)M_x^+(-v^*) \\ R_1^t(v^*) &= 0. \end{cases}$$

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

(a) Safe space equilibrium ($x = 1$). Such an equilibrium exists if and only if for some cutoff v^*

$$v^* - c + \Theta(v^*)\Delta(v^*) = h \tag{12}$$

and

$$[2F(v^*) - 1]\Theta(v^*)\Delta(v^*) \geq h. \tag{13}$$

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).[35]

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.*

(b) Transparency equilibrium ($x = 0$). Suppose now that agent i's behavior is observed by all ($x = 0$).[36] 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.

[34] We verify that $R_1^s(v^*, x)$ 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 $x = 1$.

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

[36] 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 $(-v^*, v^*)$.

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

(c) Mixed-strategy equilibrium ($0 \leq x \leq 1$). Such an equilibrium satisfies both $v^* - c + R_1^s - R_0 - h = 0$ and $R_1^s = h$. For convenience, let us assume that $\Theta(v)M^+(v)$ is weakly increasing in v. Then (using (13)), $h_2 = \Theta(c)M^+(c) \geq \Theta(v^s(h_1))M^+(v^s(h_1)) = h_1 \frac{F(v^s(h_1))}{2F(v^s(h_1))-1} > h_1$. The equilibrium is then depicted in Figure 1.

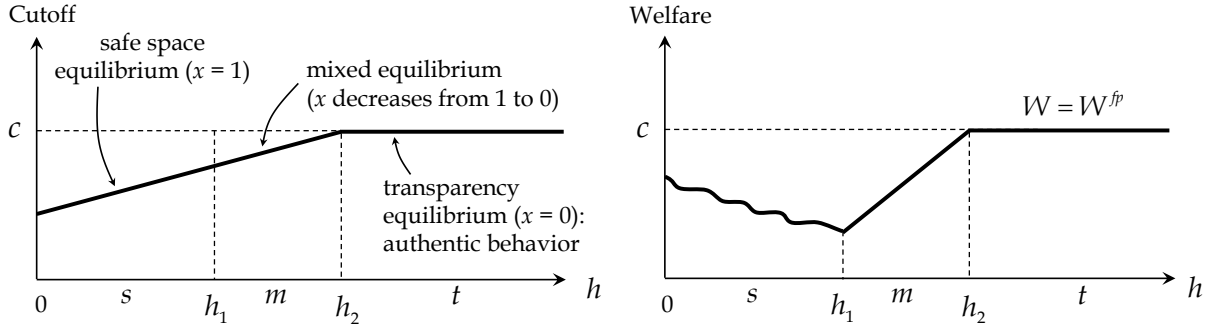

Figure 1: Cutoffs and welfare under a positional image

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:

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

Proposition 6 *(positional image). Suppose that $r(\hat{v}, v) = \mu\theta(v)\hat{v}$.*

(i) 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.

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

3.4.2 The 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 *transparency equilibrium*, the cutoff v^t is given by

$$T(v^t, 0) \equiv v^t - c - \mu[V + M^+(v^t)] + \mu V = 0 \quad \iff \quad v^t = c + \mu M^+(v^t).$$

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 transparency 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 we verify that 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, $dW^{s,m}/dh < 0$.

Proposition 7 *(maximum norm). Under the maximum norm,*

(i) The level of activity is always below the authentic level ($v^* \geq c$).

(ii) There exists $\mu_0 > 0$ such that for all $\mu \leq \mu_0$, the cutoff is continuously increasing in h, from $v^* = c$ to $v^* = v^t$. Welfare is continuously decreasing in h.

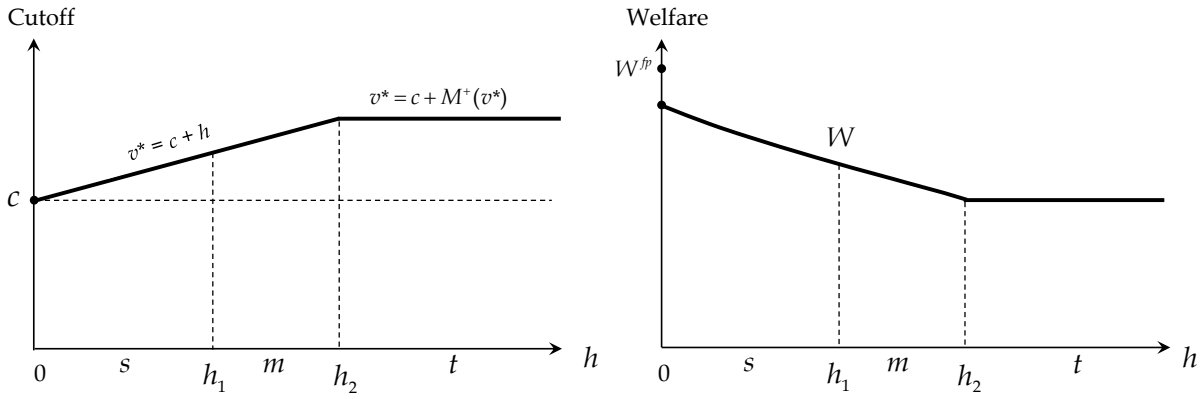

Figure 2: Cutoffs and welfare under the maximum norm

3.5 Polarization

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

Definition 3 *(polarization). Let $F(v;\rho)$ denote a smooth family of unimodal, symmetric distributions. We will say that the population as a whole 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, we focus on safe spaces (h low enough) and on a positional image. We keep making Assumption 5, which guarantees equilibrium uniqueness.

Proposition 8 *(polarization). Suppose a safe-space equilibrium (h small) and positional image concerns. A symmetric increase in polarization (in ρ) increases the fraction of active agents, $2[1 - F(v^*(\rho);\rho)]$, under a safe-space equilibrium (and, more mechanically, in a transparency equilibrium).*

Proof. The increase in the weight put in the tails ($F_\rho < 0$ for $v^* > 0$) mechanically raises activism. This is the only effect under transparency as $v^* = c$. Under a safe-space equilibrium, $v^* = v^s(\rho)$ is given by

$$v^s(\rho) - c + \Theta(v^s(\rho);\rho)\Delta(v^s(\rho);\rho) = h. \tag{14}$$

We assumed that the LHS of (14) is increasing in v^s. What about the impact of ρ? 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). ■

To further our intuition, return to the equation (14) determining the cutoff. The image incentive in a safe-space equilibrium decomposes into a *judgment-intensity parameter* $\Theta(v^s(\rho);\rho)$ corresponding to right-wing activists, and an *inference benefit* $M^+(v^s(\rho);\rho) - M^-(v^s(\rho);\rho)$ gleaned from these right-wing activists when preferring $a_i = 1$ to $a_i = 0$. In turn, we can decompose the polarization into two symmetric "one-sided polarizations" (i.e. a left-wing polarization with $F_\rho \geq 0$ for $v \leq 0$ and $F_\rho = 0$ otherwise, and a right-wing polarization with $F_\rho \leq 0$ for $v \geq 0$ and $F_\rho = 0$ otherwise). An increase in right-wing polarization boosts the right-wing safe space by making it more worthwhile to join that space by increasing both Θ and M^+. Perhaps less intuitively, the right-wing safe space, say, also expands with left-wing polarization. This is due to an heightened suspicion 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.

3.6 Asymmetric distribution

A symmetric distribution closely captures the divisiveness of the issue at stake. Asymmetric distributions are nonetheless of interest as well. Without offering a full treatment, let me make a few observations (generalizing Assumption 5 to ensure equilibrium uniqueness). First, an equilibrium is characterized by two cutoffs $\{{}^*v, v^*\}$ where ${}^*v + v^*$ in general differs from 0. Second, the equilibrium may be hybrid; for example, the majority activists may choose to be transparent while the minority activists hide in a safe space. 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 $\mathbb{R}^-$. For ε small, the issue is almost consensual. Under a positional image,[37] 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.

We now investigate the marginal impact of a small change in the distribution starting from an equilibrium characterized by cutoffs $\{{}^*v, v^*\}$; for conciseness, we again focus on a positional image and a safe-space equilibrium.[38] We look at the impact of a rise in right-wing ideology. As we will see, the impact of this evolution depends on where it is located in the type distribution:

[37] A safe space may emerge under the max norm when the issue is consensual. This does not occur with a positional image.

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

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 $v > v^*$.

Surge in acceptance of right-wing ideas. Such a surge corresponds to $F_\rho(v) \leq 0$ for $v \in ({}^*v, v^*)$ and $F_\rho(v) = 0$ otherwise.

Let us begin with the mechanical effect (composition) of an increase in ρ. 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, while $M^+(v^*;\rho)$ and $\Theta(v^*;\rho)$[39] increase. This leads to a decrease in v^*. Similarly, ${}^*v$ 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 ${}^*v$. We summarize these results in the following proposition:

Proposition 9 *(asymmetric distribution). Suppose a safe space equilibrium and positional image concerns.*

(i) A surge in right-wing extremism boosts both safe spaces.

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

3.7 The non-additive case

As discussed earlier, we want to 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 her

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

reputations vis-à-vis members of subgroups J_{-1}, J_0, J_{+1}:

$$R_i(v^*, x) \equiv \mu\Phi(R^d_{-1,a_i}(v^*, x) + R^d_{0,a_i}(v^*, x) + R^d_{1,a_i}(v^*, x)), \tag{15}$$

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 under Assumptions 1-4 for bilateral reputations:

Assumption 6 *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 3 *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.*

Lemma 3, whose proof can be found in the Appendix, shows that for image concerns that are not too important, (the counterpart of) Assumption 5 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).[40]

Proposition 10 *(non-additive case). Suppose that an agent's overall reputational payoff is given by (15), where Φ is an increasing function. Under Assumptions 5 and 6, Proposition 2 and 3 (demand for self spaces and equilibria under costless self-presentation) and Proposition 4 (equilibrium existence and characterization) hold.*

Proposition 10 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 4.

[40] For a positional image, the required Assumption 5′ guaranteeing monotonicity is satisfied if Assumption 5 is.

4 Extensions, applications and discussion

We now extend the analysis in various directions, focusing on the additive case for expositional ease (the results apply to the non-additive case as well).

4.1 The dynamics of divisive behaviors

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

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

where $R_{i,\tau}$ is i's reputational payoff at the end of date τ (for example, in a transparency equilibrium, $R_{i,\tau} = R^t(\hat{v}_{i,\tau}) \equiv \int_{-\infty}^{+\infty} r(\hat{v}_{i,\tau}, v)dF(v)$).

It turns out that the static outcome is still an equilibrium in the safe-space case, while a repeated-action outcome under transparency involves "Coasian features" and over time, puts more and more pressure on neutral types to take side.

To grasp the intuition for the *safe-space* case, 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 her behavior with her in-group, but not with her out-group, because self-presentation is costless (safe space equilibrium). Suppose that this active agent changes her behavior and becomes passive. Her former out-group does not infer anything about the change in behavior. Her former in-group observes that she defected and updates her reputation from $M^+(v^s)$ to some $\hat{v}$. Because 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 most to gain –or 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 examples in Section 2.3) of Assumption 4, 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. In words, 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). The demand for reputation then reflects

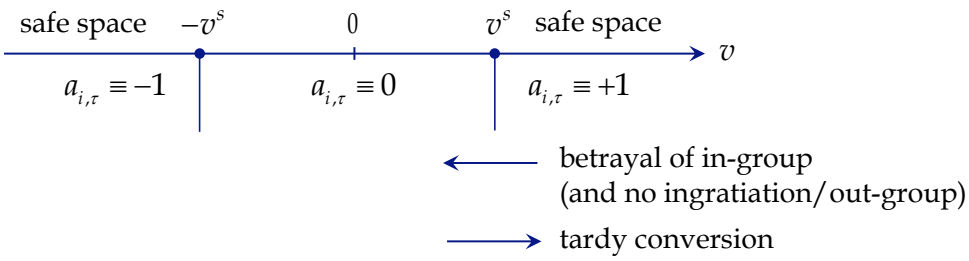

Figure 3: Low-cost self-presentation

a desire to appear moderate (Proposition 1). The agent can build a reputation for moderation by remaining passive during a few periods. The audience then "knows" that she is not an extremist and over time becomes more and more tolerant of her activism; that is, the stigma from engaging in activism is time-decreasing. 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 her 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.

Proposition 11 *(dynamics under safe spaces and transparency). Suppose that agents select actions* $\{a_{i,\tau}\}$ *sequentially at* $\tau = 0, 1, \ldots$

(i) Under safe spaces (low self-presentation cost), the static equilibrium ($a_i = 1$ *iff* $v_i \geq v^s$ *where* v^s *is given by (2) if (3) does not hold and equal to 0 otherwise) is still an equilibrium. Intuitively, defecting from the in-group is frowned upon by the latter and does nothing to ingratiate the agent with her out-group (which does not observe the change in behavior under safe spaces).*

(ii) Under transparency, the static equilibrium is no longer an equilibrium under reasonable assumptions on beliefs. By contrast, if image concerns are not too large, there exists a "Coasian equilibrium" in which the fraction of active agents increases over time.

Figure 4 illustrates part (ii) of Proposition 11 in the case of the maximum norm and continuous time. The Appendix derives the necessary conditions for a Coasian equilibrium in the general transparency case under discrete time.

4.2 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.[41] In practice, outings tend to be more frequent for high-image-

[41] We are 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

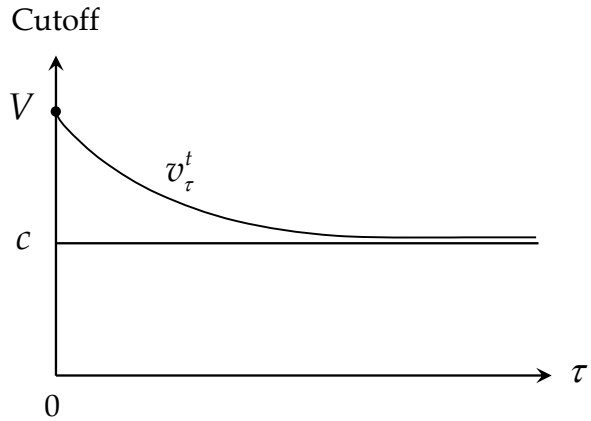

Figure 4: Dynamics under high hiding cost, the maximum norm and continuous time

concerns members (politicians, celebrities, local notables...). While the basic theory developed so far predicts why such members are hurt more by the outing, it does 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, I 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 Proposition 3(ii)). 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.

The Appendix develops a simple version of this argument. It shows in particular 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).

4.3 Endogenous social graphs and ghettoisation

A complementary reason why revealing one's behavior to the in-group only is costly is that one may have to direct one's social graph (friends, colleagues, club mates) toward

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".

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 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. We capture this in a stark form: Agent i, when choosing $a_i = +1$ benefits from a safe space if and only if her 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 a loss of opportunities (friends are selected in a smaller group, leading to a lower average match quality) and/or a lack of diversity (if diversity is valued in and of itself). Furthermore, it may involve a transition cost of making new friends if one already has a diversified circle of friends.

To capture this second, *switching cost*, let us assume that each agent has a fixed number of friends and that there is a unit cost of making new friends and another unit cost of abandoning old ones. Intuitively, friends comprise only a small proportion of the overall population. We keep a continuum for expositional simplicity, but this implies nothing as to the relative masses of friends and overall audience. Starting with a diversified sets of friends and moving from density $f(v)$ to density $g(v)$, the cost of making new friends is proportional to the number of new friends $\int_{-\infty}^{+\infty}(g-f)^+dv$ and the cost of abandoning friends proportional to $\int_{-\infty}^{+\infty}[-(g-f)^-]dv$.[42] But $\int_{-\infty}^{+\infty}(g-f)^+dv = -\int_{-\infty}^{+\infty}(g-f)^-dv$, so in the end the sum of the two costs is proportional to the L^1-norm

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

One can capture the first cost, associated with the *lack of diversity*, in a similar way through the distance between the potential social graph and the selected one. Let the "natural" social graph of "first-best" matching opportunities be given by density $f(v)$ (i.e. the best matching opportunities reflect the overall population): It provides the agent with the maximal choice for her social graph. Other "second-best" opportunities come at a unit cost. The number of lost social opportunities or second-best friends is equal to $\int_{-\infty}^{+\infty}(g-f)^+dv = \|g-f\|/2$. The choice of social graph $g(v)$, with which information about one's behavior will be mechanically shared, again gives rise to a cost proportional to the L^1-norm distance $\|f-g\|$ between the two distributions.

In either case, the agent i picking $a_i = +1$ really has the choice between two disclosure strategies: Be transparent and then select the optimal social graph f, or select friends in $[v^*, +\infty)$ with $v^* \geq 0$ (say, by joining the corresponding safe space) to avoid leaks about her behavior, yielding social graph $g(v) = f(v)/[1-F(v^*)]$ on $[v^*, +\infty)$ and a cost of preserving privacy equal to say, $\kappa\|f-g\|/2$, or

$$h(v^*) = \frac{\kappa}{2}\left[\int_{-\infty}^{v^*}|f(v)-0|dv + \int_{v^*}^{+\infty}\left|f(v) - \frac{f(v)}{1-F(v^*)}\right|dv\right] = \kappa F(v^*).$$

[42] $(g(v)-f(v))^+ \equiv \max\{0, g(v)-f(v)\}$ and $(g(v)-f(v))^- \equiv \min\{0, g(v)-f(v)\}$.

The hiding cost grows as fewer agents act (as v^* grows). A *safe space* equilibrium $v^* = v^s$ satisfies

$$S(v^s, 1) = 2\kappa F(v^s) \quad \text{and} \quad R_1^s(v^s, 1) - R_1^t(v^s, 1) \geq \kappa F(v^s)$$

where, as earlier,

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

A *transparency* equilibrium $v^* = v^t$ does not require changing friends or limiting the diversity of the circle of friends ($h \equiv 0$) and satisfies

$$T(v^t, 0) = 0 \quad \text{and} \quad R_1^s(v^t, 0) - R_1^t(v^t, 0) \leq \kappa F(v^t)$$

where, as earlier:

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

The endogenous hiding cost interestingly is a factor of *strategic complementarity.* Consider a transparent equilibrium; the cutoff $v^* = v^t$ is high (greater than c). So retrenching into a safe space involves a high cost in terms of diversity/lost opportunities ($F(v^*)$ is high). Conversely, in a safe space equilibrium, v^* is low (below c) and so the cost of joining a safe space is lower. As a consequence, there may be multiple equilibria even if Assumption 5 holds.

The intuition goes as follows: 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 changing friends is lower. That makes the cost of securing privacy lower, making safe spaces more attractive.

One can indeed find examples in which (a) the equilibrium is unique when the hiding cost h is exogenous, but (b) there are multiple equilibria when $h = \kappa F(v^*)$.

Dynamics of the social graph. Let us now index periods by $\tau \in \{0, 1 \ldots\}$. We must 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; let h_τ denote the date-τ hiding cost and g_τ the social graph of the agent. Then the date-τ cost for agent i is

$$h_{i,t} \equiv \kappa_\delta \|f - g_{i,t}\| + \kappa_\sigma \|g_{i,t} - g_{i,t-1}\|.$$

The first cost diversity cost (κ_δ) is recurrent; the second cost (κ_σ) corresponds to switching in the social graph. Once the agent has changed friends to accommodate her 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 increasing the intensity of image concerns μ. The mixing of the two communities may be permanently undone even as the identity would turn less salient again ceteris paribus.

Proposition 12 *(endogenous hiding costs). When the hiding cost is generated by a lack of diversity or a cost of switching acquaintances,*

(i) Its endogeneity $(h(v^*))$ *is a factor of strategic complementarity* $(h'(v^*) > 0)$.

(ii) The individual social graph exhibits hysteresis.

4.4 Reputation as a random member of a group

We have so far assumed that the representative member of her perceived group defines an agent's reputation. That is, the reputational payoff of an agent i vis-à-vis an agent j with 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 her 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 Assumptions 1-5.

Proposition 13 *(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 her 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, transparency) equilibrium exists if and only if* $h \leq h_1$ *(resp.* $h \in [h_1, h_2]$, $h \geq h_2$*) for some* $h_2 > h_1 > 0$.

5 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 really wish; this section focuses on this latter effect. So far, signaling occurs 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 members of the outgroup. There are two possible rationales for this.

a) Signaling to the community. Our first reason why within-in-group signaling will occur is that agents want to show they are "the true believers".[43]

As an illustration in Facebook groups and other fora inhabited by like-minded individuals, *one-sided information and narratives* circulate within the group (for signaling reasons: no-one wants to be perceived as spreaders of group-adverse messages). 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 are voluntary, even though they 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 would not voluntarily be chosen by members but serve the community as a whole.

Let us illustrate the first motive, signaling to the community, for instance. Suppose that $v \sim [-V, +V]$ and that $r(\hat{v}, v) \equiv \mu\theta(v)\hat{v}$ (positional reputations). At cost c, an individual can engage in normal/minimal compliance in activity $|a_i| = 1$. He can also show zeal and select $z \geq 0$ (for "zeal") at cost

$$c + |V - v|z + \frac{z^2}{2}, \tag{16}$$

where $c < V$. So, even leaving aside negative externalities on the rest of society, zeal is wasteful from the point of view of the community.

Under transparency a_i and (if $|a_i| = 1$) z_i are observed by all. In a safe space, a_i and z_i are observed only within the community. Suppose for illustrative purposes that $\Theta(v^*) \equiv \mu \int_{v^*}^{V} \theta(v)dF(v)$. The following proposition is proved in the online Appendix.

Proposition 14 *(one-upmanship). Assume positional image concerns and a cost of acting with or without zeal given by (16).*

(i) For h low enough, there exists a symmetric safe space equilibrium in which (for $v_i \geq 0$, and symmetrically for $v_i \leq 0$):

[43]This is amplified when the cost of joining a safe space involves changing one's social graph; the cost is then a joint cost as it applies to other signaling activities. The group then descends in one-upmanship.

- *types* $v_i \leq v^*$ *do not act* ($a_i = 0$),
- *types* $v_i \in [v^*, \tilde{v}]$ *act without zeal* ($z = 0$),
- *types* $v_i \in [\tilde{v}, V]$ *act with zeal* $z(v) = \Theta(v^*) - (V - v)$.

(ii) In contrast, for a high hiding cost, the equilibrium is transparent and the equilibrium zeal is $z(v) \equiv 0$ *for all* z. *The equilibrium cutoff is* $v^* = c$.

The two dark sides of safe spaces (voluntary and 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 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), I identify these collateral effects of the formation of safe spaces as the very reason why such communities turn into "tribes".

Remark (splinter groups). The escalation of partisanship studied in Proposition 14 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 wants to, which need not be the case) by excluding extremists, depriving the latter from an audience. To illustrate this point in the simplest possible way, assume that the agent's type v is distributed on $[-V, +V]$, that reputation is positional, and that there is no hiding cost so that a safe-space equilibrium prevails. Now augment the action space to include two elements on each side, say on the right side $a = 1$ and $A > a$. So an activist can behave as a simple militant (intrinsic motivation va) or as a radical (intrinsic motivation vA). The safe space equilibrium in which all activists pick the moderate action a (or $-a$) is still an equilibrium if $V(A - a) \leq \Theta(v^*)\Delta(v^*)$. The RHS of this inequality captures the idea that becoming more radical and leaving the safe space does not alter the opinion of the safe space's outgroup, but does so for the in-group: the agent is perceived by safe-space members as part of the outgroup rather than the in-group by safe-space members.[44]

6 Alleys 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 is a bit of a neglected field. The paper developed a conceptual framework to study such environments. When people do not agree on what's right or wrong, transparency lead some to alter their behavior or to take refuge in a safe space. The paper then 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), create reputational externalities on moderates (who are suspected by

[44] When $V = +\infty$, then some always become radical, but they may be a tiny minority (interestingly, such radicalism then exhibits strategic complementarities).

both sides and pushed to pick one), and generate tribalistic over-signaling beyond desired practice (either voluntarily, or coerced through the threat of outing). We also saw how symmetric increases in polarization lead to an increase in tribalism, 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 ghettoization is dynamically stable and even subject to hysteresis. Rather than describing more in detail the results (this was done in the introduction), I would like to conclude for a few (of the many) alleys for future research.

Needless to say, the paper touched on only a subset of questions related to the public and private spheres and to authenticity. For example, technology does not only mechanistically expand the public sphere. It also offers new opportunities for interactions, which may make existing relationships more transient as people seize these opportunities. One might conjecture that reputational concerns might decline together with the expected length of relationships; yet this is not so, as platform business models have amply demonstrated: The new technologies not only enable people to get in contact and to introduce themselves or their goods, but they also generate a demand for, and allow a greater transparency through mutual ratings, and thereby strong reputational concerns. Visibility is again endogenous.

Relatedly, people can seek transparency or, to the contrary, take refuge in anonymity. Privacy protects us from the need to "posture" to improve our social image. As we have seen, if hostile opinions weigh more than favorable ones, privacy allows agents to undertake activities they like, but are controversial. But we also need a private sphere for reasons that are not captured by the model: we may want to let off steam without antagonizing others, especially if our frustration has multiple causes or we had a bad day that alters our judgment; we want to think aloud, throw ideas around and share them with others, that may not be the right ones or might even offend some; we may want to share thoughts and feelings with like-minded peers without hurting others. It may even be that privacy helps us abide by our duty to respect others as our public discourse is then better guided by reflection.

The useful 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. Social norms within a like-minded group or imposed by a majoritarian group in the population 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.[45]

The search for private spheres may take the form of the creation of a fake identity (catfishing) or more subtly the belonging to groups of like-minded peers on social networks or in physical spaces (like in gated communities, or the rural or urban communities where

[45]Similarly, political correctness (Morris 2001), like the notion of a safe space, 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.

some actors of the 1968 contest movement took refuge from an oppressive society). This quest for an authenticity enabled by a smaller need for posturing 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.[46] The contours of the private and public spheres are not set only by technology, they are also socially determined by explicit individual choices. This feature represents a fruitful direction for future research.

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.

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. We mentioned in the introduction 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.

[46]This is related to the abuse of the concept of a "safe space", understood by some as a space that is expunged of individuals with conflicting opinions (see e.g. Lukianoff-Haidt 2019).

References

Acemoglu, D., Makhdoumi, A., Malekian, A. and A. Ozdaglar (2022), "Too Much Data: Prices and Inefficiencies in Data Markets," forthcoming*American Economic Journal: Microeconomics*.

Acquisti, A., Taylor, C., and L. Wagman (2016), "The Economics of Privacy," *Journal of Economic Literature*, 54: 442–492.

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

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

Algan, Y., Dalvit, N., Do, Q.A., Le Chapelain, A., and Y. Zenou (2019), "Friendship Networks and Political Opinions: A Natural Experiment Among Future French Politicians", mimeo.

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

Allport, G. W. (1954), *The Nature of Prejudice*, Cambridge, MA: Perseus Books.

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.

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. (2022), "Scoring Strategic Agents", mimeo.

Banks, J., and J. Sobel (1987), "Equilibrium Selection in Signaling Games," *Econometrica*, 55(3): 647–661. DOI: 10.2307/1913604.

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

Bénabou, R., and J. Tirole (2011b), "Laws and Norms", NBER WP 17579.

Bénabou, R. and J. Tirole (2016), "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, 30(3): 141–164.

Bénabou, R., Falk, A. and J. Tirole (2020), "Narratives, Imperatives and Moral Reasoning," mimeo.

Bergemann, D., Bonatti, A. and T. Gan (2022), "The Economics of Social Data," *Rand Journal of Economics*, 53(2): 263–296.

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

Bernheim, D., and A. Bodoh-Creed (2023), "Pervasive Signaling", *Theoretical Economics*, 18: 163–196.

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

Bordalo, P., Tabellini, M., and D. Yang (2022), "Issue Salience and Political Stereotypes," mimeo.

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

Braghieri, L. (2021), "Political Correctness, Social Image, and Information Transmission," R&R, *American Economic Review.*

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

Choi, J.P, Jeon, D.S, and B. C. Kim (2019), "Privacy and Personal Data Collection with Information Externalities," *Journal of Public Economics*, 173: 113–124.

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.

Dewatripont, M., Jewitt, I. and J. Tirole (1999), "The Economics of Career Concerns, I: Comparing Information Structure," *Review of Economic Studies*, 66(1): 183–198.

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.

Fudenberg, D., Levine, D. and J. Tirole (1985) *Infinite-Horizon Models of Bargaining with One-Sided Incomplete Information*, in A. Roth (ed.), Game Theoretic Models of Bargaining, Cambridge University Press, 73-98.

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.

Golub, B., and M. Jackson (2012), "How Homophily Affects the Speed of Learning and Best-Response Dynamics," *Quarterly Journal of Economics*, 127(3): 1287–1338.

Gul, F., Sonnenschein, H. and R. Wilson (1986), "Foundations of Dynamic Monopoly and the Coase Conjecture," *Journal of Economic Theory*, 39: 155–190.

Habermas, J. (1962), *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*, Polity, Cambridge (trans. 1989).

Harbaugh, R., and E. Rasmusen (2018), "Coarse Grades: Informing the Public by Withholding Information," *American Economic Journal: Microeconomics*, 10: 210–23.

Hirshleifer, J. (1971), "The Private and Social Value of Information and the Reward to Inventive Activity," *American Economic Review*, 61: 561–574.

Holmström, B., and P. Milgrom (1991), "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 7: 24–52.

Hong, F., Tirole, J. and C. Zhang (2023), "Prosocial Behavior in Public and Private Spheres: Theory and Evidence," 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. (2021), "Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone," mimeo, UC Berkeley.

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.

Lanier, J. (2010), "Missing Persons," In *You Are Not a Gadget,* London: Penguin.

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

Lewis, T., and D. Sappington (1989), "Countervailing Incentives in Agency Problems," *Journal of Economic Theory*, 49: 294–313.

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

Maggi, G., and A. Rodriguez-Clare (1995), "On Countervailing Incentives," *Journal of Economic Theory*, 66: 238–263.

Manski, C., and J. Mayshar (2003), "Private Incentives and Social Interactions: Fertility Puzzles in Israel," *Journal of the European Economic Association*, (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.

Milgrom, P. and S. Oster (1987), "Job Discrimination, Market Forces, and the Invisibility Hypothesis," *Quarterly Journal of Economics*, 102(3): 453–476.

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.

Pesendorfer, W. (1995), "Design Innovation and Fashion Cycles," *American Economic Review*, 85: 771–792.

Posner, R.A. (1978), "The Right of Privacy," *Georgia Law Review*, 12(3): 393–422.

Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, Penguin Random House.

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

Smith, A. (1759), *The Theory of Moral Sentiments*, Strand & Edinburgh: A. Millar.

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

Tirole, J. (2021), "Digital Dystopia," *American Economic Review*, 111(6): 2007–2048.

Veblen, T. (1899), "The Theory of the Leisure Class: An Economic Study in the Evolution of Institutions," New York: Macmillan.

Williams, B. (1985), *Ethics and the Limits of Philosophy*, Routledge.

Appendix

Proof of Lemma 1 (modified L^p norm satisfies Assumptions 1-4)

We actually prove a stronger form of Assumption 4: 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 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

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

Proof of non-existence of asymmetric equilibria (Proposition 4)

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 $\hat{v}$.

We need to generalize Assumption 5 to the asymmetric-behavior case. Let

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

We add to Assumption 5:

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 J_0 agents. For x^* we have:

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

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

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

where the last equality is due to Assumption 1. 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 5 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 2 implies that $x^* = {}^*x$.

Hence, we only need to check the case in which $x^* = 1$ and ${}^*x$ belongs to $(0,1)$. Assumption 2, 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:

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

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:

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

When J_1 is a transparent group ($x^* = 0$), then:

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

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 1 entail that the equation giving ${}^*v$ is:

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

Assumption 2, and the fact that $M(-v^*, -{}^*v) > M^-(-{}^*v)$, imply that:

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

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

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

using Assumption 1, and the symmetry of the distribution of types yields:

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

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

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

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 2 implies that ${}^*x \leq 0$, a contradiction.

■

Proof of Proposition 5 (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 5.

(ii) Recall that $W = R_0 + 2\int_{v^*}^{+\infty}(v-v^*)dF(v)$, 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,

$$R_0^{fp} = \int_{-\infty}^{+\infty} r(0,v)dF(v) = R_0^t \geq \max\{R_0^s, R_0^m\}.$$

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 $x = 0$. Assumption 2 then implies that

$$R_0^{fp} = \int_{-\infty}^{+\infty} r(0, v) dF(v) \geq \int_{-\infty}^{+\infty} r(\hat{v}_0(v), v) dF(v).$$

■

Equilibrium existence under a positional image

Let

$$\begin{cases} R_1^s(v^*, x) \text{ denote the weighted image when choosing } a_i = 1 \text{ and hiding it from non-peers} \\ R_1^t(v^*) \text{ denote the weighted image when choosing } a_i = 1 \text{ and being transparent} \\ R_0(v^*, x) \text{ denote the weighted image when choosing } a_i = 0. \end{cases}$$

Suppose that individuals who act (say, $a_i = 1$) hide with probability x and remain transparent with probability $1 - x$.

Then

$$\begin{cases} R_1^s(v^*, x) & = \Theta(v^*) M^+(v^*) - \Theta(v^*) M^+(v^*) \dfrac{x[1 - F(v^*)]}{x[1 - F(v^*)] + [2F(v^*) - 1]} \\ R_1^t(v^*) & = 0 \\ R_0(v^*, x) & = -2\Theta(v^*) M^+(v^*) \dfrac{x[1 - F(v^*)]}{x[1 - F(v^*)] + [2F(v^*) - 1]} \end{cases}$$

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

$$v^* - c + R_1^s(v^*, x) - R_0(v^*, x) = h \quad \Leftrightarrow \quad v^* - c + \Theta(v^*) \left[\frac{2x[1 - F(v^*)] + 2F(v^*) - 1}{x[1 - F(v^*)] + 2F(v^*) - 1} \right] M^+(v^*) = h \tag{17}$$

$$R_1^s(v^*, x) - R_1^t(v^*) = h \quad \Leftrightarrow \quad v^* - c = -2\Theta(v^*) \left[\frac{x[1 - F(v^*)]}{x[1 - F(v^*)] + [2F(v^*) - 1]} \right] M^+(v^*) \tag{18}$$

The redundant condition implied by (17) and (18), is type $v \geq v^*$'s indifference between transparency and safe space when $a = 1$:

$$\Theta(v^*) \left[\frac{2F(v^*) - 1}{x[1 - F(v^*)] + [2F(v^*) - 1]} \right] M^+(v^*) = h. \tag{19}$$

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 (12), 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 7 *(Assumption 6 applied to positional image).* $c > 2\Theta(0)M^+(0)$.

Conditions (17) and (18) 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

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

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

Proof of Lemma 3 (Assumption 5 is satisfied for the true L^p norm)

Let us show that Assumption 5 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

$$\begin{cases} K_T(v^*, x) = -\dfrac{1}{\mu}\big[R_1^t(v^*) - R_0(v^*, x)\big] \\ \\ K_S(v^*, x) = -\dfrac{1}{\mu}\big[R_1^s(v^*, x) - R_0(v^*, x)\big], \end{cases}$$

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.

$$R_1^t(v^*) = -\mu \left[\int_{-\infty}^{M^+(v^*)} (M^+(v^*) - v)^p dF(v) + \int_{M^+(v^*)}^{+\infty} (v - M^+(v^*))^p dF(v) \right]^{\frac{1}{p}}.$$

Let:

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

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

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:

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

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.

$$-\frac{1}{\mu}R_0(v^*,x) \;=\; \left[2\Big(\int_{v^*}^{+\infty}(-M_x^-(v^*)+v)^p dF(v) + \int_0^{v^*} v^p dF(v)\Big)\right]^{\frac{1}{p}}.$$

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

$$\frac{\partial(-\frac{1}{\mu}R_0(v^*,x))}{\partial v^*} \equiv \frac{N_1+N_2}{D}$$
$$= \frac{\overbrace{\frac{2}{p}f(v^*)(v^{*p}-(v^*-M_x^-(v^*))^p}^{N_1} + \overbrace{2(-M_x^-(v^*))'\int_{v^*}^{+\infty}(-M_x^-(v^*)+v)^{p-1}dF(v)}^{N_2}}{\underbrace{\left[2\Big(\int_{v^*}^{+\infty}(-M_x^-(v^*)+v)^p dF(v) + \int_0^{v^*} v^p dF(v)\Big)\right]^{\frac{p-1}{p}}}_{D}}.$$

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

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

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 $|\frac{-1}{\mu}\frac{\partial R_1^s(v^*,x)}{\partial v^*}| < M$ for all $x \in [0,1]$ and $v^* > V$.

$$\begin{aligned}
-\frac{1}{\mu}R_1^s(v^*,x) &= \Big[\int_{v^*}^{M^+(v^*)}(M^+(v^*)-v)^p dF(v) + \int_{M^+(v^*)}^{+\infty}(v-M^+(v^*))^p dF(v) \\
&+ 2\int_0^{v^*} v^p dF(v) + \int_{-\infty}^{-v^*}(M_x^+(-v^*)-v)^p dF(v)\Big]^{\frac{1}{p}}.
\end{aligned}$$

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

$$\frac{\partial(-\frac{1}{\mu}R_1^s(v^*,x))}{\partial v^*} \equiv \frac{N}{D},$$

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

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

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

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

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

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

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

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

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 $v^* > V$. ■

Proof of Proposition 11 (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.[47] 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 $v^t - c + R^t(v^t) = R^t(0)$):

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

[47] 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 her 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 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.

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

$$v_k^t - c + R(M(v_k^t, v_{k+1}^t)) = (1-\delta)R(0) + \delta[v_k^t - c + R(M(v_{k-1}^t, v_k^t))] \tag{20}$$

and

$$v_0^t - c + R(M(v_0^t, v_1^t)) = R(0), \tag{21}$$

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 (20) 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

$$\frac{d\tau}{dv} = \frac{R'(v)}{i(v-c)}. \tag{22}$$

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

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

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

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

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

More general assumptions

Assumptions 1 through 3 provide micro-foundations for more general, reduced form assumptions that drive the preceding results. Let $R_J(\hat{v})$ denote the reputational payoff of an agent with reputation $\hat{v}$ within audience subgroup J. In a symmetric equilibrium with

cutoff $v^* \geq 0$ and hiding probability x, $J \in \{J_{-1}, J_0, J_{+1}\}$ where $J_{-1} \equiv (-\infty, -v^*)$, $J_0 \equiv (-v^*, v^*)$, $J_1 \equiv (v^*, +\infty)$.

Assumption 1′ (symmetry). For all $\hat{v}$,

$$R_{J_1}(\hat{v}) = R_{J_{-1}}(-\hat{v}) \text{ and } R_{J_0}(\hat{v}) = R_{J_0}(-\hat{v}).$$

Assumption 2′ (distate for dissonance). Suppose that $-v^* < \hat{v} < M^+(v^*)$. Then

$$R_{J_{-1}}(\hat{v}) > R_{J_{-1}}(M^+(v^*)).$$

Assumption 3′ (benefit from being perceived as representative of a group rather than the marginal type at the lower end). Whenever $\hat{v} < v^*$,

$$R_{J_1}(M^+(v^*)) \geq R_{J_1}(v^*) \geq R_{J_1}(\hat{v})$$

Assumption 4′ (concavity). For all $\hat{v}$

$$R_{J_0}(0) \geq R_{J_0}(\hat{v}).$$

For equilibrium behavior $\{v^*, x\}$, let

$$\begin{aligned}
R_1^s(v^*, x) &\equiv [1 - F(v^*)]R_{J_1}(M^+(v^*)) + [F(v^*) - F(-v^*)]R_{J_0}(0) \\
&\quad + F(-v^*)R_{J_{-1}}(M_x^+(-v^*)) \\
R_1^t(v^*) &\equiv F(v^*)R_{J_{-1}}(M^+(v^*)) + [F(v^*) - F(-v^*)]R_{J_0}(M^+(v^*)) \\
&\quad + [1 - F(v^*)]R_{J_1}(M^+(v^*)) \\
R_0(v^*, x) &\equiv [F(v^*) - F(-v^*)]R_{J_0}(0) + 2[1 - F(v^*)]R_{J_1}(M_x^-(v^*)),
\end{aligned}$$

denote the reputational payoffs attached to $a_i = +1$, when the agent operates in a safe space ($R_1^s(v^*, x)$) or opts for transparency ($R_1^t(v^*)$), and attached to being neutral ($R_0(v^*, x)$). Assumption 1 through 3 on bilateral reputations r imply Assumption 1′ through 4′ on overall reputations (R_1^s, R_1^t, R_0).

The results of this paper more generally are obtained using Assumptions 1′ through 4′, as well as Assumption 5, which is already expressed in terms of overall reputations rather than bilateral ones.

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 $\mu(1-\alpha)\alpha zw$ 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

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

Euclidean image concerns

Suppose now that $r(\hat{v}, v) = -\mu(\hat{v} - v)^2$. Reputations are[48]

[48] And so,

$$R_1^s(v^*, x) - R_0(v^*, x) = \mu[1 - F(v^*)][M^+(v^*) - M_x^-(v^*)]^2 = \mu[1 - F(v^*)](1+B)^2[M^+(v^*)]^2$$

and

$$R_0(v^*, x) - R_1^t(v^*) = \mu[1 - 2[1 - F(v^*)][B^2 + BA]][M^+(v^*)]^2,$$

where

$$B \equiv \frac{x[1 - F(v^*)]}{x[1 - F(v^*)] + [2F(v^*) - 1]}.$$

Let $v^*(h)$ denote a root of the following equation:

$$S(v^*, x) \equiv v^* - c + [R_1^s(v^*, x) - R_0(v^*, x)] = h.$$

In particular $v^*(0) < c$.

$$\begin{cases} R_1^s(v^*,x) &= -\mu[\sigma^2 + [1-F(v^*)][-(M^+(v^*))^2 + (M_x^-(v^*))^2 - 2M_x^-(v^*)M^+(v^*)]] \\ R_1^t(v^*) &= -\mu[\sigma^2 + (M^+(v^*))^2] \\ R_0(v^*,x) &= -\mu[\sigma^2 + 2[1-F(v^*)][(M_x^-(v^*))^2 - 2M^+(v^*)M_x^-(v^*)]] \end{cases}$$

Safe space region. Assumption 6, which guarantees that $v^*(0) > 0$ writes in the quadratic case: $c > 2\mu[M^+(0)]^2$. A safe space equilibrium satisfies, if interior ($v^* > 0$):

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

or

$$v^* - c + \mu[1 - F(v^*)][M^+(v^*) - M^-(v^*)]^2 = h.$$

Using the identity $[1-F(v^*)]M^+(v^*)+F(v^*)M^-(v^*) = E[v] = 0$, this condition, together with the condition that disclosure is not optimal, can be rewritten as

$$v^* - c + \mu\frac{1-F(v^*)}{(F(v^*))^2}(M^+(v^*))^2 = h \leq \mu\left[\frac{1-F(v^*)+F^2(v^*)}{F^2(v^*)}\right](M^+(v^*))^2. \tag{23}$$

(Analysis of mixed equilibrium to be performed)

Transparency region. When the agents' behaviors are observable by all, the reputational payoff when choosing $a_i = 0$ is $E[-\mu v^2] = -\mu\sigma^2$, letting σ^2 denote the variance of v. The reputational payoff when choosing $a_i = 1$ is $E[-\mu[v - M^+(v^*)]^2] = -\mu\sigma^2 - \mu(M^+(v^*))^2$.

And so the cutoff $v^* > c$ is given by

$$v^* - c - \mu(M^+(v^*))^2 = 0 \geq \mu[2 - F(v^*)](M^+(v^*))^2 - h. \tag{24}$$

If the support of F is infinite, $M^+(v^*) \geq v^*$ implies that

$$T(v^*, 0) \equiv v^* - c - \mu(M^+(v^*))^2$$

goes to $-\infty$ as v^* goes to $+\infty$. Also $T(c) < 0$. Assume that T is concave in v^*, which is indeed the case in the following examples:

- *Uniform distribution:* $v \sim U[-V, +V]$. Then $M^+(v^*) = \frac{v^*+V}{2}$.
- *Exponential distribution:* $1 - F(v) = e^{-\lambda v}/2$ for $v \geq 0$ (the distribution is mirrored for $v < 0$). Then $M^+(v^*) = v^* + \lambda^{-1}$.
- *Pareto-distribution:* $1 - F(v) = (v_0/v)^p/2$ for $v \geq v_0 > 0$ (again the distribution is mirrored for $v < 0$) and $p > 1$. Then $M^+(v^*) = pv^*/(p-1)$.

More generally, the monotone hazard rate condition ($f/[1-F]$ increasing) is a sufficient condition for $T(v^*, 0)$ to be concave:

$$\begin{aligned}\frac{d^2}{dv^2}(M^+(v^*))^2 &= 2\Big[\Big(\frac{f(v^*)}{1-F(v^*)}\Big)^2[(M^+(v^*)-v^*)^2 + M^+(v^*)(M^+(v^*)-v^*)] \\ &+ M^+(v^*)[M^+(v^*)-v^*]\frac{d}{dv^*}\Big(\frac{f(v^*)}{1-F(v^*)}\Big)\Big].\end{aligned}$$

For distributions with an infinite support such as the exponential and the Pareto distributions, the properties that $S(c) < 0$, $S(+\infty) < 0$ and the concavity of S imply that there are 0 or 2 solutions to (24). If image concerns are not too large ($\mu \leq \bar{\mu}$ for some $\bar{\mu}$), then there are two solutions. Only the smaller of the two solutions is stable. When $\mu > \bar{\mu}$, the solution is $v^* = +\infty$: $a_i = 0$ for all v.

To do. Can one have multiple equilibria (say, one safe space, one transparent) with Euclidean image concerns?

Proof of Proposition 13 (reputation as a random member of a group)

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

Assumption 4′ *(benefit from being perceived by the in-group as representative of the in-group rather than as a passive type (J_0) in the population).*

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

Assumption 4′ is satisfied for a positional image. [CHECK THAT IT IS SATISFIED FOR THE MODIFIED L^p NORM]

We first check that Proposition 1 (on the demand for reputation) holds: Consider a symmetric equilibrium $\{v^*, x\}$. And let

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

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

denote the reputational payoff of a member of group 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 her action is transparent, and, when joining a safe space,

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

using Assumption 2 ($r(\tilde{v}, v)$ is decreasing in $\tilde{v}$ for $v \leq -v^* \leq \tilde{v}$).

Similarly, agent i's reputational payoff vis-à-vis agent j with $a_j = 0$ is $\Gamma^1(v, v^*)$ if her 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 Proposition 2 (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 2, $\int_{-v^*}^{v^*} r(\tilde{v}, v) dF(v)$ is concave in $\tilde{v}$, and peaks at 0. Hence we have:

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

On the other hand, we know that:

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

where we invoke again Lemma 2: $\int_{-v^*}^{v^*} r(\tilde{v}, v) dF(v)$ 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:

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

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′.

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

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

using Assumption 2 ($r(\tilde{v}, v)$ 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 Proposition 4. 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):

$$\begin{cases} R_1^s(v^*, x) & \equiv & \int_{v^*}^{+\infty} \Gamma^1(v, v^*) dF(v) dF(v) + \int_{-v^*}^{v^*} \Gamma_x^0(v, -v^*) dF(v) + \int_{-\infty}^{-v^*} \Gamma_x^1(v, -v^*) dF(v) dF(v) \\ R_1^t(v^*) & = & \int_{-\infty}^{+\infty} \Gamma^1(v, v^*) dF(v) \\ R_0(v^*, x) & \equiv & \int_{v^*}^{+\infty} \Gamma_x^1(-v, -v^*) dF(v) + \int_{-v^*}^{v^*} \Gamma_x^0(v, -v^*) dF(v) + \int_{-\infty}^{-v^*} \Gamma_x^1(v, -v^*) dF(v). \end{cases}$$

Now, define:

$$S(v^*, x) \equiv v^* - c + R_1^s(v^*, x) - R_0(v^*, x)$$
$$= v^* - c + \int_{v^*}^{+\infty} \Big[\Gamma^1(v, v^*) - \Gamma_x^1(-v, -v^*) \Big] dF(v)$$

denote the net benefit from acting in a safe spaces and

$$T(v^*, x) \equiv v^* - c + R_1^t(v^*) - R_0(v^*, x),$$

denote the net benefit from acting transparently. The rest of the proof is entirely similar to the proof of Proposition 4 for the existence and characterization of the symmetric equilibrium. ■

Proof of Proposition 14 (one-upmanship)

Safe spaces equilibrium. Suppose there is no cost of hiding when $|a_i| = 1$. First, we look for a separating equilibrium on $[\tilde{v}, V]$. Suppose that, within the community choosing $a_i = 1$, the reputation grows one-for-one with z whenever $z > 0$: $d\hat{v}/dz = 1$.

Fix v^*. Conditionally on choosing $a_i = 1$, type v solves

$$u(v) \equiv \max_z \left\{ v - c - (V - v)z - \frac{z^2}{2} + \Theta(v^*)\hat{v}(z) \right\}$$

yielding (for z positive)

$$z(v) = \Theta(v^*) - (V - v),$$

verifying our hypothesis. Letting $M(v^*, \tilde{v}) = E[v|v \in [v^*, \tilde{v}]]$, pooling at $z = 0$ yields payoff:

$$u_0(v, \tilde{v}) = v - c + \Theta(v^*)M(v^*, \tilde{v}).$$

Note that $u(V) - u_0(V, V) = \Theta(v^*)[V - M^+(v^*)] - \frac{\Theta^2(v^*)}{2} > 0$ if μ is not too large. Conversely $u(v^*) - u_0(v^*, v^*) = -(V - v^*)z(v^*) - \frac{z^2(v^*)}{2} < 0$. So there exists a cutoff $\tilde{v}$ such that $u(\tilde{v}) = u_0(\tilde{v}, \tilde{v})$. Zeal jumps from 0 to $\Theta(v^*) - (V - \tilde{v})$ at $\tilde{v}$. This cutoff is a function $\tilde{v}(v^*)$ of v^*.

The cutoff v^* is then given by

$$v^* - c + \Theta(v^*)[M(v^*, \tilde{v}(v^*)) - M^-(v^*)] = 0.$$

Transparency. In this case, the choice (a_i, z_i) is observed by all and so $\int_{-V}^{+V} \mu\theta(v)\hat{v}(a_i, z_i) = 0$, implying authentic behavior. ■