PNAS RESEARCH ARTICLE | POLITICAL SCIENCES

Strategic analysis of dissent and self-censorship

Joshua J. Daymude[a,1,2], Robert Axelrod[b,1,2], and Stephanie Forrest[a,c,1]

Contributed by Robert Axelrod; received April 4, 2025; accepted September 10, 2025; reviewed by Margaret Roberts and Bruce Schneier

Expressions of dissent against authority are an important feature of most societies, and efforts to suppress such expressions are common. Modern digital communications, social media, and Internet surveillance and censorship technologies are changing the landscape of public speech and dissent. Especially in authoritarian settings, individuals must assess the risk of voicing their true opinions or choose self-censorship, voluntarily moderating their behavior to comply with authority. We present a model in which individuals strategically manage the tradeoff between expressing dissent and avoiding punishment through self-censorship while an authority adapts its policies to minimize both total expressed dissent and punishment costs. We study the model analytically and in simulation to derive conditions separating defiant individuals who express their desired dissent in spite of punishment from self-censoring individuals who fully or partially limit their expression. We find that for any population, there exists an authority policy that leads to total self-censorship. However, the probability and time for an initially moderate, locally adaptive authority to suppress dissent depend critically on the population's willingness to withstand punishment early on, which can deter the authority from adopting more extreme policies.

dissent | civil resistance | self-censorship | surveillance | strategic decision-making

Significance

Expressions of dissent against authority are an important feature of most societies while efforts to suppress such expressions are common. With the development of social media, expressions of dissent are easy to circulate. Meanwhile, authorities have gained powerful surveillance technologies for identifying objectionable messages and their authors, using these to deter or punish observed dissent. We present a model of individuals and an authority who adapt to each other, showing the conditions under which individuals choose to censor themselves and how authorities balance their antipathy to dissent against punishment costs. The results predict that an adaptive authority can always suppress dissent but that the time to do so depends superlinearly on the boldness of the population.

Dissent against authority is ubiquitous throughout human history and culture (1), as are authorities' efforts to maintain power, stability, and social order. But modern Internet infrastructure, encrypted communication protocols, social media platforms, and automated surveillance technologies pose new opportunities and threats for potential dissidents (2, 3). Fueled by the last decade of advances in computer vision (4), facial recognition tools in particular have been used around the world to track and arrest protesters, inducing a chilling effect (5–10). Internet censorship systems such as China's Great Firewall (11–13) and Russia's "TSPU" devices for centralized, nation-wide, real-time deep packet inspection (14, 15) monitor traffic for site access or material considered objectionable. The Cyberspace Administration of China explicitly signals its ability to punish what it finds: "The cyber police are right by your side... You will exercise restraint and rationality when you post and write messages" (16). The United States values free speech as a constitutional right; however, there are numerous tensions around privacy and acceptable speech, ranging from U.S. technology giants' heterogeneous and opaque practices of "content moderation" (17–27) to the FBI's four million warrantless data queries of American's "incidentally collected" private communications in 2020–2023 (28) to the January–February 2025 Executive Orders aiming to "end federal censorship" while simultaneously deeming any "diversity, equity, and inclusion" language and programs within the federal government illegal (29–31).

In these settings, individuals must make careful choices about if, when, and how to express dissent: Do they voice their true opinions in spite of the potential consequences, or do they engage in self-censorship (32–35)—voluntarily moderating their behavior to avoid punishment by the authority—thus granting the authority an implicit form of control? These considerations point to a set of strategic questions about self-censorship: When it emerges, when it is resisted, its relationship to different punishment regimes, and how interactions between the population and the authority change these dynamics over time.

This work resides in the gap between the intuitive sense and formative theory of civil resistance as a distributed, strategic, and uncertain process (36–38) and the population-level, comparative, statistical methods typically used to study them (39, 40). We present a model comprising a single authority and a population of individuals. The individuals strategically decide what level of dissent to express based on the strength of their inherent desire to dissent, their boldness, and the perceived risk of punishment by the authority. We derive analytical results for the individuals' optimal actions under two forms of authoritarian punishment, uniform and proportional, and compare the model's predictions to contemporary examples. Finally, we investigate how a population's desired

Author affiliations: [a]Biodesign Center for Biocomputing, Security and Society and the School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281; [b]School of Public Policy, University of Michigan, Ann Arbor, MI 48109; and [c]Santa Fe Institute, Santa Fe, NM 87501

Author contributions: J.J.D., R.A., and S.F. devised the mathematical model; J.J.D. performed all mathematical analysis, implemented and conducted the simulation experiments, and produced visualizations; and J.J.D., R.A., and S.F. interpreted the results and wrote the paper.

Reviewers: M.R., University of California San Diego; and B.S., Harvard University.

The authors declare no competing interest.

Copyright © 2025 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]J.J.D., R.A., and S.F. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: jdaymude@asu.edu or axe@umich.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2508028122/-/DCSupplemental.

Published November 5, 2025.

Downloaded from https://www.pnas.org by Casey Luskin on January 13, 2026 from IP address 72.132.200.2.

dissent and boldness distributions influence a locally adaptive authority's policies, and how that in turn affects population behavior.

Model

We model a single authority A (e.g., a government, corporation, or social media platform) governing a population of n individuals $\mathcal{I} = \{I_1, \ldots, I_n\}$ (e.g., citizens, employees, or users, respectively). Each individual I_i has a level of desired dissent denoted by $\delta_i \in [0, 1]$, where $\delta_i = 0$ is no desire and $\delta_i = 1$ is maximal. In each round $r = 1, 2, \ldots$, each individual I_i takes an *action* with a level of expressed dissent denoted by $a_{i,r} \in [0, \delta_i]$. The authority then observes each action with probability

$$\Pr[A \text{ observes } a_{i,r}] = \nu_r + (1 - \nu_r) \cdot a_{i,r}, \qquad [1]$$

where $\nu_r \in [0, 1]$ is the authority's *surveillance* in round r. This reflects an assumption that the more dissenting an action is, the more likely the authority is to observe it. The authority then punishes each individual I_i according to its *punishment function* π based on the actions it observes, its *tolerance* $\tau_r \in [0, 1]$, and its *severity* $\psi_r > 0$ in round r:

$$\pi(a_{i,r}, \tau_r, \psi_r, \nu_r) = \begin{cases} 0 & \text{if } A \text{ does not observe } a_{i,r} \text{ or } a_{i,r} \le \tau_r; \\ \psi_r & \text{if } A \text{ observes } a_{i,r} > \tau_r \text{ and } \pi \text{ is uniform;} \\ \psi_r \cdot (a_{i,r} - \tau_r) & \text{if } A \text{ observes } a_{i,r} > \tau_r \text{ and } \pi \text{ is proportional.} \end{cases} \qquad [2]$$

We consider two functional forms for π: a *uniform* function where the authority punishes all infractions equally and a *proportional* (linear) function where "the punishment fits the crime." For example, a government restricting press access for any reporter it views as antagonistic is uniform punishment, while fining a speeding driver based on their infraction over the speed limit is proportional punishment. In digital spaces, an example of uniform punishment is Internet shutdowns (i.e., blocking all Internet access regardless of content) while social media companies' content moderation policies often employ proportional punishment, with consequences ranging from content removal to account demonetization to permanent bans depending on the type and frequency of violations (41). Only actions that exceed the authority's tolerance τ_r and are observed by the authority are punished.

The authority's utility at the end of round r reflects its preference for *i*) as little dissent as possible while *ii*) incurring as little punishment cost as possible:

$$U_{A,r} = -\alpha \cdot \sum_{i=1}^{n} a_{i,r} - \sum_{i=1}^{n} \pi(a_{i,r}, \tau_r, \psi_r, \nu_r), \qquad [3]$$

where $\alpha > 0$ is the authority's *adamancy*, which reflects the relative importance of expressed dissent compared to the cost of punishing. The utility of an individual I_i at the end of round r represents its preference for *i*) acting as closely as possible to its desire while *ii*) avoiding punishment:

$$U_{i,r} = \beta_i \cdot (1 - \delta_i + a_{i,r}) - \pi(a_{i,r}, \tau_r, \psi_r, \nu_r), \qquad [4]$$

where $\beta_i > 0$ is a constant representing how *bold* individual I_i is, i.e., the relative strength of its preference to act according to its desires relative to the need to avoid punishment.

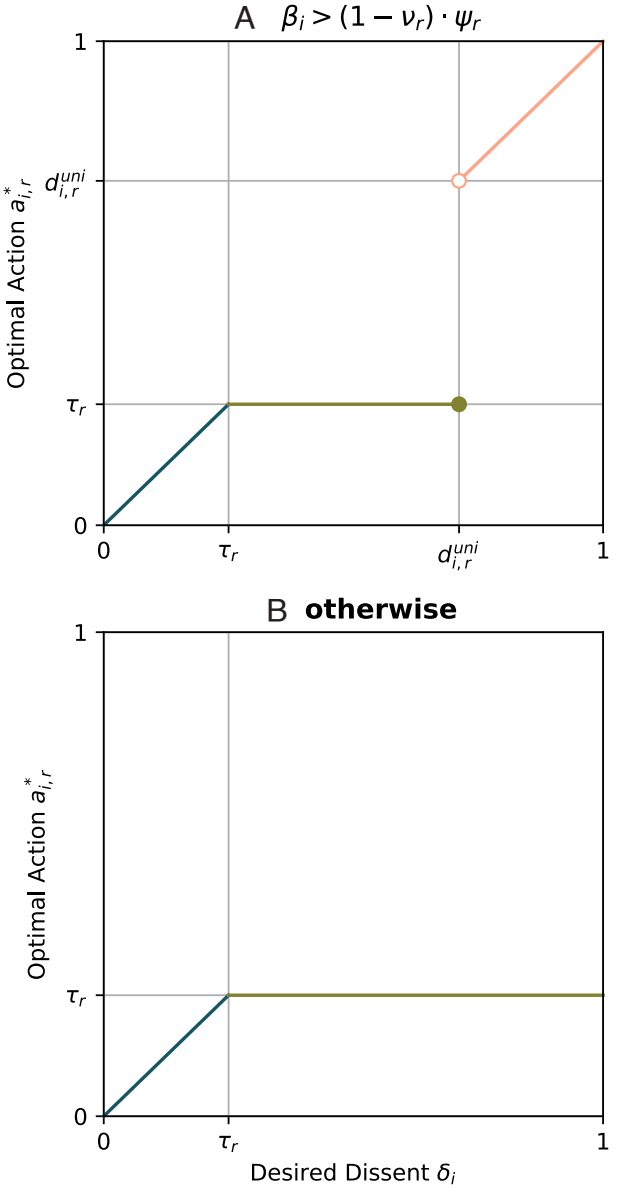


Fig. 1. Individuals' optimal actions under uniform punishment. The optimal action of an individual I_i as a function of its desired dissent δ_i when π is uniform (Eq. **5**) and (*A*) sufficient boldness $\beta_i > (1 - \nu_r) \cdot \psi_r$ makes defiance possible or (*B*) otherwise, when full self-censorship always occurs above τ_r. Compliance is shown in teal, self-censorship in olive, and defiance in peach.

Results

Individuals' Optimal Actions: Compliance, Self-Censorship, and Defiance. We begin by analyzing and interpreting the optimal (utility-maximizing) action $a^*_{i,r}$ for an individual I_i in round r when the authority's parameters are fixed. Beyond its own desired dissent δ_i and boldness β_i, we assume that individual I_i knows the authority's punishment function π and the values of the authority's tolerance τ_r, severity ψ_r, and surveillance ν_r.* Complete derivations of all results in this section are given in *SI Appendix*.

Uniform punishment. When π is uniform, the optimal action for an individual I_i is

*All results in this section also hold if the authority's exact parameter values are replaced by individuals' noisy estimates thereof; see *SI Appendix* for details.

Downloaded from https://www.pnas.org by Casey Luskin on January 13, 2026 from IP address 72.132.200.2.

$$a^*_{i,r} = \begin{cases} \delta_i & \text{if } \delta_i \le \tau_r \\ & \text{or } \left(\beta_i > (1-\nu_r)\cdot\psi_r \text{ and } \delta_i > d^{\text{uni}}_{i,r}\right); \\ \tau_r & \text{otherwise,} \end{cases} \quad [5]$$

where

$$d^{\text{uni}}_{i,r} = \frac{\beta_i \cdot \tau_r + \nu_r \cdot \psi_r}{\beta_i - (1-\nu_r)\cdot\psi_r}. \quad [6]$$

Fig. 1 shows $a^*_{i,r}$ as a function of desired dissent δ_i.

This result categorizes individual responses into three types according to their desired dissents: *i*) *compliant* individuals whose desired dissents are below the authority's tolerance, enabling them to act as they prefer without punishment; *ii*) *self-censoring* individuals whose desired dissent exceeds the authority's tolerance but who act at the authority's tolerance level instead, preferring to avoid punishment to expressing dissent; and *iii*) *defiant* individuals whose boldness and desired dissents are so large that the utility lost by not acting as they prefer outweighs the cost of punishment. Of these three types, only self-censoring individuals act at lower levels of dissent than their desires.

The most interesting result of this analysis is the distinction between self-censorship and defiance. Individuals in both categories desire to act above the authority's tolerance, but analysis of Eqs. **5** and **6** shows that defiance requires $\beta_i > \psi_r/(1-\tau_r)$; i.e., any defiant individual must be at least as bold as the severity of punishment.

Proportional punishment. When π is proportional to the level of expressed dissent, the optimal action for an individual I_i is

$$a^*_{i,r} = \begin{cases} \delta_i & \text{if } \delta_i \le \tau_r \text{ or } (\nu_r = 1 \text{ and } \beta_i \ge \psi_r) \\ & \text{or } (\nu_r < 1 \text{ and } \delta_i \le d^{\text{pro}}_{i,r}); \\ \tau_r & \text{if } \delta_i > \tau_r \text{ and}((\nu_r = 1 \text{ and } \beta_i < \psi_r) \\ & \text{or } (\nu_r < 1 \text{ and } \tau_r \ge d^{\text{pro}}_{i,r})); \\ d^{\text{pro}}_{i,r} & \text{otherwise,} \end{cases} \quad [7]$$

where

$$d^{\text{pro}}_{i,r} = \frac{1}{2}\left(\tau_r + \frac{\beta_i - \nu_r\cdot\psi_r}{(1-\nu_r)\cdot\psi_r}\right). \quad [8]$$

As in the case when π is uniform, there are *compliant* individuals whose preferred actions are below the authority's tolerance, allowing them to act as desired without punishment. But surprisingly, in the regime of imperfect surveillance, proportional punishment inverts the parameter regions where which self-censorship and defiance occur (Fig. 2*A*). Here, *defiant* individuals who act as they desire despite potential punishment are not those with the largest desired dissent—as they are when π is uniform—but those with intermediate desires. Intuitively, because both the probability of being observed and the subsequent punishment if observed scale linearly with the infraction, small infractions can be worth the relatively low risk of detection, if it means acting as desired. However, above the tipping point where the risk of detection and cost of subsequent punishment outweigh desire, individuals become *partially self-censoring*, acting below their desired dissent but still above the authority's tolerance. The existence and precise location of the tipping point depends critically on the relation between boldness, tolerance, severity, and surveillance. This partial self-censorship arises only under proportional punishment.

If the authority's surveillance is perfect ($\nu_r = 1$), observing every infraction, then an individual's optimal action depends only on the relation between its boldness and the authority's severity: If $\beta_i \ge \psi_r$, then I_i is defiant, acting exactly as it desires regardless of punishment (Fig. 2*B*); otherwise, it self-censors to avoid punishment (Fig. 2*C*). Thus, perfect surveillance removes the optimal action's dependence on desired dissent: Defiant individuals are sufficiently bold; fully self-censoring individuals are not.

Contemporary Illustrations of the Model. Contemporary examples provide natural illustrations of the phases of behavior predicted by our model (Fig. 3). As a simple example, after evidence of NSA/PRISM online surveillance was widely publicized in June 2013 (i.e., recognition of increased surveillance without a corresponding change in tolerance or severity), Internet traffic to Wikipedia articles on privacy-sensitive topics significantly and immediately decreased (42), demonstrating a self-censoring chilling effect consistent with our model's predictions (Fig. 3*B*). Survey results from 2017 by the same author found that personal legal threats (i.e., credible threats of high-severity punishment) and guarantees of government surveillance drove greater rates of self-censorship than relevant statues communicating the authority's tolerance (43), aligning qualitatively with the roles of these parameters in Eqs. **6** and **8** (Fig. 3 *A*–*C*).

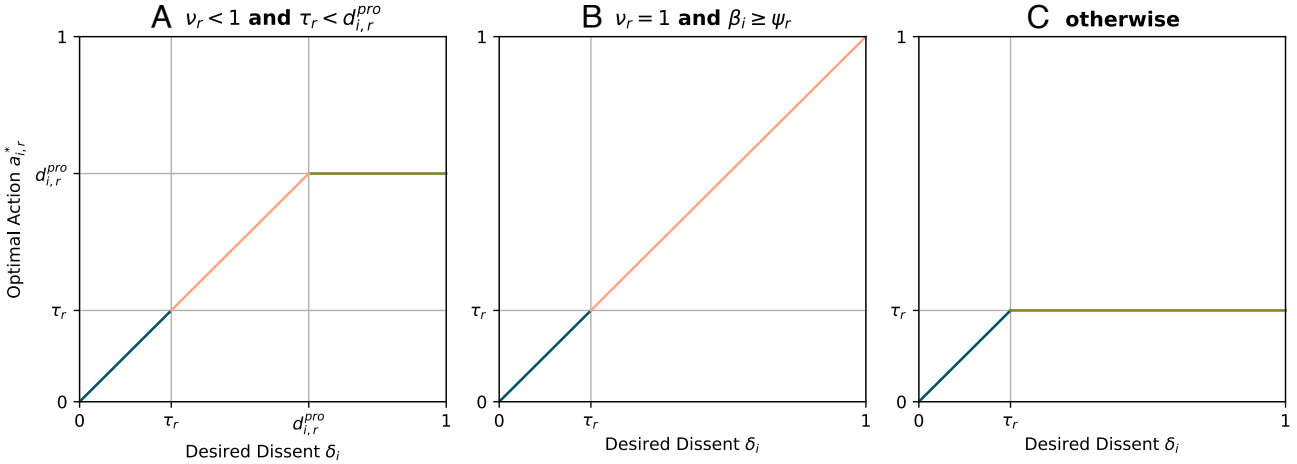


Fig. 2. Individuals' optimal actions under proportional punishment. The optimal action of an individual I_i as a function of its desired dissent δ_i when π is proportional (Eq. **7**) and (*A*) the authority's surveillance is imperfect with $\nu_r < 1$ and the authority's tolerance $\tau_r < d^{\text{pro}}_{i,r}$ is insufficient to eliminate defiance, (*B*) the authority's surveillance is perfect with $\nu_r = 1$ but the individual's boldness $\beta_i \ge \psi_r$ yields only defiance above τ_r, or (*C*) otherwise, when full self-censorship always occurs above τ_r. Compliance is shown in teal, self-censorship in olive, and defiance in peach.

Downloaded from https://www.pnas.org by Casey Luskin on January 13, 2026 from IP address 72.132.200.2.

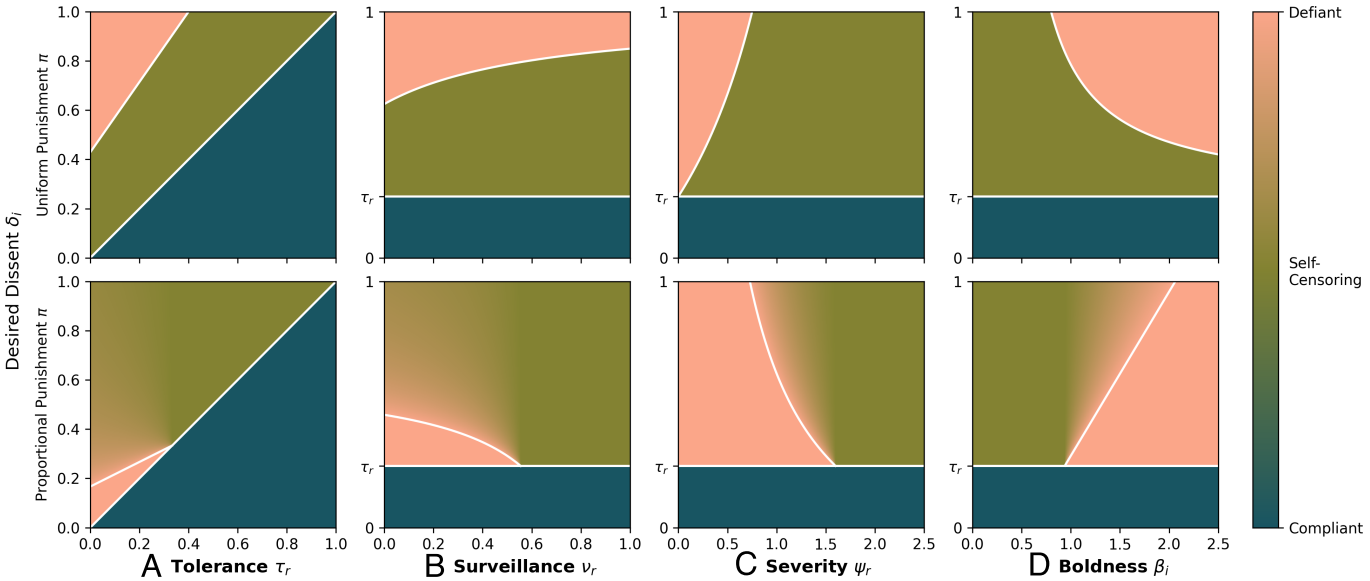


Fig. 3. Phase diagram of individuals' optimal behaviors by parameter. Compliant (teal), self-censoring (olive), and defiant (peach) phases of individuals' optimal actions are shown as a function of (*A*) the authority's tolerance τ_r, (*B*) the authority's surveillance ν_r, (*C*) the authority's severity ψ_r, and (*D*) the individual's boldness β_i vs. their desired dissent δ_i for both uniform (*Top* row) and proportional (*Bottom* row) punishment. The phase boundaries delineating defiance (peach) from self-censorship (olive) occur at critical dissent values $d_{i,r}^{\text{uni}}$ (Eq. **6**, *Top* row) and $d_{i,r}^{\text{pro}}$ (Eq. **8**, *Bottom* row). The specific parameter values used to obtain this visualization were chosen to show all possible phases of behavior in each pairwise sweep, but they are representative of a wider range of parameter values.

Both China and Russia can be assumed to have low tolerance of dissent and high-severity proportional punishments. Further, both regimes have made significant investments in their digital infrastructure to achieve self-censorship, as illustrated by the famous analogy of the "anaconda in the chandelier" (44) and China's 2019 ranking as the most digitally repressive country in the world (3). In these settings, our model predicts that only low levels of dissent are commonly expressed, as small infractions are both less likely to be observed and have low expected punishment cost (Fig. 2*A*). For example, the popular 2021 meme of "Lying Flat" in China expressed disaffection with hustle culture and contemporary globalization, but was largely confined to online spaces where the stakes were low (45). This is reminiscent of the dictator's dilemma (46) which suggests that authoritarian repression leads to distorted information feedback, creating space for low-risk dissent to flourish. However, Russia's recently deployed TSPU (14, 15) has even deeper surveillance and more granular censorship controls than China's Great Firewall and related censorship technologies (11–13). Our model predicts that, holding other parameters constant, Russia's increased surveillance capabilities will drive greater rates of self-censorship than in China (Fig. 3*B*).

In the social media space where surveillance is high, a study that measured self-censorship on Twitter during the Hong Kong protests of 2022 to 2023 (47) found that after the passage of a new national security law (i.e., a decrease in tolerance and increase in severity), Hong Kong users self-censored by deleting or restricting their accounts, removing old posts, and moving away from politically sensitive topics (Fig. 3 *A* and *C*). In the opposite direction (increased tolerance and no severity), users of Gab.com accelerated rates of hate speech after establishing that Gab's content moderation policies tolerated such behavior and did not threaten any serious consequences (48). Similarly, Meta's January 2025 reversal of content moderation policies is an example of simultaneously increasing tolerance and removing all severity (27).

An Adaptive Authority. The authority's utility minimizes both the population's total expression of dissent (i.e., its *political cost*) and its *punishment cost* (Eq. **3**). Perhaps unsurprisingly, there always exists a draconian policy for the authority to suppress all dissent through self-censorship without having to enact any punishments. Formally, given any population's desired dissent and boldness parameters $\{(\delta_i, \beta_i)\}_{i=1}^n$, one utility-maximizing policy for the authority is

$$\tau^* = 0,$$
$$\nu^* = 1, \text{ and} \qquad [9]$$
$$\psi^* > \begin{cases} \max_i\{\delta_i \cdot \beta_i\} & \text{if } \pi \text{ is uniform;} \\ \max_i\{\beta_i\} & \text{if } \pi \text{ is proportional.} \end{cases}$$

In words, this draconian policy tolerates nothing, observes everything, and threatens to punish so extremely that all individuals completely self-censor, expressing no dissent at all. This policy is, in fact, a Nash equilibrium for the leader–follower game in which the authority acts first and the individuals act second, without collusion and under the assumption that the authority's threat of punishment is fully credible. The lower bounds on severity ψ^* result from Eqs. **5**–**8** when $\tau_r = 0$ and $\nu_r = 1$; see *SI Appendix* for a derivation. Notably, the draconian severity ψ^* for uniform punishment is at most that of proportional punishment.

In reality, it is unlikely that an authority knows all individuals' desired dissent and boldness parameters a priori or exists in a political context where such an extreme policy could be introduced immediately. Thus, we also consider how an initially moderate authority might incrementally adapt its policies over time to increase its utility, using feedback through the political and punishment costs incurred in each round to determine how to change its policy. A round is the time it takes for the authority to observe dissent and enact a new policy, which we take to be roughly days to weeks.

We frame the authority's policy adaptation as the local greedy process of random mutation hill climbing (49). The authority is

Downloaded from https://www.pnas.org by Casey Luskin on January 13, 2026 from IP address 72.132.200.2.

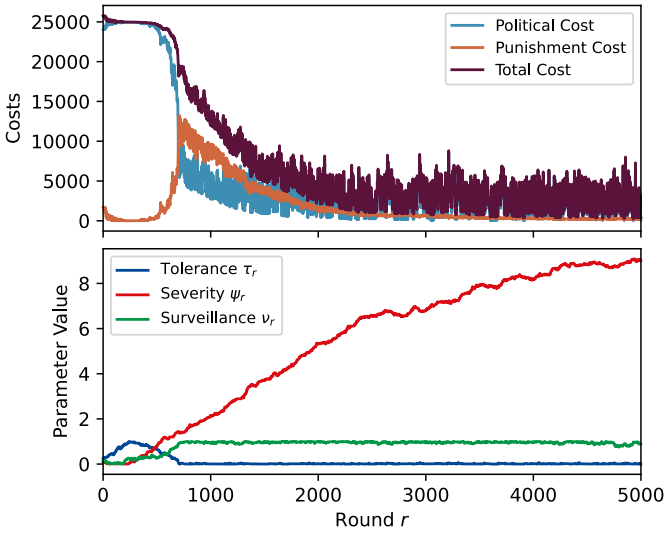


Fig. 4. Time evolution of an adaptive authority. The adaptive authority's costs (*Top*) and parameters (*Bottom*) over 5,000 rounds with uniform punishment and adamancy $\alpha = 1$. The population has $n = 100{,}000$ individuals whose desired dissents are sampled independently from a truncated exponential distribution with mean 0.25 and bounds [0, 1]. Likewise, their boldness constants are sampled independently from an exponential distributions with mean 2.0. Recall that the authority's tolerance τ_r and surveillance ν_r (*Bottom*, blue and green, resp.) are bounded in $[0, 1]$ while its severity ψ_r (*Bottom*, red) can be any positive value. Corresponding results for proportional punishment can be found in *SI Appendix*, Fig. S1.

initialized with a moderate policy (τ_0, ψ_0, ν_0) chosen uniformly at random in $[0, 1]^3$. Subsequently, in each round, it chooses one of its three parameters at random, mutates that parameter to a new value within an ϵ-radius of its original value, and adopts that as the new policy. If the new policy decreases the authority's total cost, it is retained; otherwise, the mutation is reverted (see *SI Appendix*, Algorithm 1 for details). This simple rule captures incremental policy change by an initially moderate authority, ignoring any exogenous shocks.

Fig. 4 shows an example run of this adaptive authority's evolving policies and corresponding costs over time. Initially ($r = 0$), the authority has low tolerance (Fig. 4, *Bottom*, blue). With insufficient severity and surveillance to back that preference (Fig. 4, *Bottom*, red and green), the authority incurs high political cost as the entire population expresses its desired dissent (Fig. 4, *Top*, blue). Perhaps counterintuitively, the authority does not respond by immediately raising its severity and surveillance, as this would saddle it with the costly task of punishing the vast majority of the population. Instead, it keeps severity and surveillance low while gradually increasing tolerance until all dissent is allowed (Fig. 4, *Bottom*, blue shows $\tau_r = 1$ near round $r \approx 250$). The authority then reverses course ($250 \leq r \leq 750$), gradually decreasing its tolerance while increasing its severity and surveillance. This causes a cascade of self-censorship: High-dissent individuals start by giving up just a small amount of expressed dissent to stay within the authority's tolerance, but then capitulate further and further to match the authority's decreasing tolerance. As severity and surveillance increase simultaneously, it becomes too costly to regain defiance. Individuals who are sufficiently bold remain defiant through the cascade, causing an initial spike in the authority's punishment costs (Fig. 4, *Top*, orange, $r \approx 750$), but even they are progressively persuaded to self-censor by continuously increasing punishment severity and high surveillance (Fig. 4, *Bottom*, red and green, $r \geq 1{,}000$). By round $r = 5{,}000$, the authority approximates the draconian policy, achieving near-total suppression of dissent. Similar patterns occur across wide parameter ranges (Fig. 5) and under both punishment functions (*SI Appendix*, Fig. S1).

The sequence shown in Fig. 4 is reminiscent of Mao's 1956–1957 "Hundred Flowers Campaign" in which outspoken dissident elites were encouraged to freely criticize the government for the dual purpose of relieving tensions and strengthening the socialist system (50). After a brief period of escalating critique that culminated in popular demands for power sharing (51)—i.e., large political costs resulting from high tolerance—Mao reversed course with the Anti-Rightist Movement, sending thousands of outspoken dissidents to penal camps and labor exile—i.e., a policy of little-to-no tolerance and high severity—dissuading those who remained from expressing further dissent.

Although a draconian policy always exists, the adaptive authority's probability of discovering one depends on the population's desired dissent and boldness distributions (Fig. 5). In particular, Fig. 5*C* shows a curved boundary delineating a high-boldness region where the authority incurs large total costs from a lower-boldness region of minimal total costs. In the latter, the authority has discovered a near-draconian policy that neither tolerates dissent (Fig. 5 *A* and *D*, lower region) nor enacts much punishment (Fig. 5 *B*, *Lower* region); instead, with high surveillance and severity that increases with population boldness (Fig. 5 *E* and *F*, *Lower* region), it achieves mass self-censorship. On the other hand, when the population's mean desired dissent is moderate ($\delta \in [0.05, 0.35]$) and mean boldness β is sufficiently high, the authority almost never discovers a

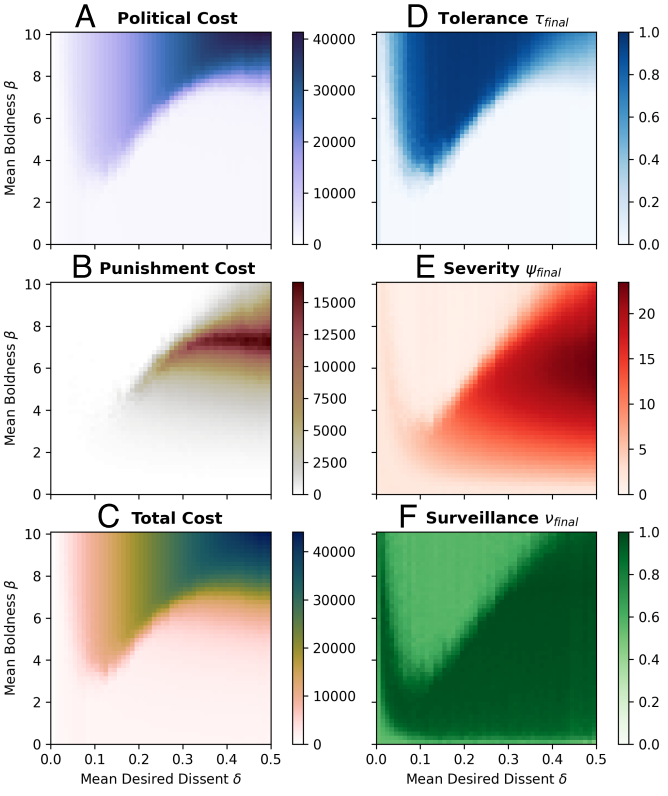


Fig. 5. Authority costs and policies after adaptation under uniform punishment. The authority's average final (*A*–*C*) costs and (*D*–*F*) parameter values after 10,000 rounds of random mutation hill climbing adaptation across 50 independent trials per (δ, β) pair, where $\delta \in [0.005, 0.495]$ and $\beta \in [0.1, 10]$ are the means of the population's exponential distributions of desired dissents and boldness constants, respectively. Each trial involves $n = 100{,}000$ individuals and the authority uses uniform punishment π with adamancy $\alpha = 1$. Corresponding results for proportional punishment can be found in *SI Appendix*, Fig. S2.

Downloaded from https://www.pnas.org by Casey Luskin on January 13, 2026 from IP address 72.132.200.2.

suppressing policy (Fig. 5 *C*, *Upper* region). Instead, the authority tolerates essentially all dissent (Fig. 5 *A* and *D*, *Upper* region), reducing severity and making surveillance irrelevant (Fig. 5 *E* and *F*, *Upper* region). This bifurcation of suppressing vs. tolerating policies also occurs under proportional punishment (*SI Appendix*, Fig. S2), with a critical mean boldness β^* separating the two.

Between these stable outcomes of suppression through self-censorship and tolerance of all dissent is a small region of realized punishment for moderate-high desired dissent, moderate-boldness populations (Fig. 5*B*, maroon region). This is not a distinct third outcome, but rather a snapshot of transient dynamics in which an authority has successfully driven low-desire/boldness individuals to self-censor but must punish numerous high-desire/boldness individuals who defiantly express dissent (similar to Fig. 4, $r \approx 750$). Even though the authority will eventually achieve near-total suppression of these populations, the time it takes to do so depends primarily on the boldness of the population. Thus, while no dissent movement can last forever against increasingly severe authoritarian punishment, increasing a population's mean boldness has a superlinear effect on the length of time that population can sustain dissent and attempt to enact change (*SI Appendix*, Fig. S3).

Discussion

Individuals aim to express a desired level of dissent while avoiding punishment. This leads them to be compliant, self-censoring, or defiant depending on their and the authority's parameters (Figs. 1 and 2). Under uniform punishment, when all infractions are punished identically, self-censorship occurs when individuals have an intermediate level of desired dissent, while a sufficiently high desire leads to defiance (Fig. 1*A*). Surprisingly, under proportional punishment, the pattern is reversed, with moderate individuals acting as desired while extreme individuals partially self-censor (Fig. 2*A*). In both settings, a critical value of desired dissent separates self-censorship from defiance (Eqs. **6** and **8**). As a fraction of the population, self-censorship increases as the authority's tolerance, severity, or surveillance increases, or as individual boldness decreases (Fig. 3). There is always a draconian policy with which the authority achieves total self-censorship (Eq. **9**). However, when an initially moderate authority adapts incrementally, the population's mean boldness—i.e., individuals' willingness to express dissent in spite of punishment—determines whether the authority suppresses dissent and how long it takes to do so (Figs. 4 and 5 and *SI Appendix*, Figs. S1–S3).

Neither an authority's ability to observe dissent nor self-censorship are new phenomena. For example, leading a mass gathering or publishing a critical article with a real byline were always visible actions, and many countries' laws prohibit sedition and blasphemy and support substantial self-censorship. But modern surveillance technologies have blurred the lines between public and private actions, potentially exposing even minor expressions of dissent to scrutiny and punishment. As ubiquitous surveillance becomes less an issue of technical capability than one of willingness, norms, and regulations (3, 4, 8, 15), it will be increasingly challenging to preserve a culture of free expression (Figs. 3*B* and 4).

The model shows that the most powerful lever a population has for sustaining dissent is its boldness, i.e., the strength of individuals' preferences to act according to their desired dissent. Under uniform punishment, high boldness is a necessary condition for any individual to express dissent (Eq. **5**); under proportional punishment and perfect surveillance ($\nu_r = 1$), the boldness-to-severity ratio completely determines whether defiance or self-censorship occurs (Fig. 2 *B* and *C*). More importantly, the larger a population's mean boldness, the more slowly an initially moderate, locally adaptive authority discovers a policy that suppresses dissent, if it ever does (Fig. 5 and *SI Appendix*, Fig. S3). For example, early and widespread public backlash to privacy-invasive technologies such as China's Green Dam Youth Escort system in 2009 (52) and the recent Microsoft Recall feature in Windows 11 (53, 54) significantly postponed both projects, even though both eventually reemerged in modified forms. Threat of punishment may encourage individuals to self-censor, but our results suggest that an adaptive authority does not enact large-scale punishment until after the masses have become compliant (Fig. 4). It is simply too costly for the authority to punish everyone, but once the vast majority of individuals become compliant or self-censoring, it is cost-effective for the authority to enact increasingly severe punishments against remaining dissidents. This is reminiscent of the often-quoted first lesson of Snyder's *On Tyranny*: "Do not obey in advance" (55). Preemptive surrender through self-censorship, before punishment is imposed, is a fast path to authoritarian control. Thus, a population's persistent dissent in spite of threatened punishment early on may deter an authority from adopting more extreme policies.

Promising directions for future work include extending the analyses and simulations to include individuals that adapt their desired level of dissent and boldness values over time. These parameter adaptations could be strategic in response to the authority's changing policies, or they could arise from opinion dynamics with individuals arranged in networks of social influence akin to Penney's theory of chilling effects as social conformity (56). Our preliminary results (*SI Appendix*) suggest that assimilative opinion dynamics alone—whether on desired dissent, boldness, or both—are insufficient for a small minority of high-desire, high-boldness dissidents to drive a larger cascade of defiance akin to Chenoweth and Stephan's empirical "3.5% Rule" (39, 40). Further investigation could characterize more carefully which behaviors and parameters are most salient in determining when a networked population behaves differently from the setting we considered here.

Materials and Methods

Derivations of individuals' optimal actions under uniform and proportional punishment start from an individual's utility function and analyze its maximization with respect to the authority's parameters. The existence of a Nash equilibrium at the authority's draconian policy follows from similar arguments; see *SI Appendix* for details. Pseudocode detailing the random mutation hill climbing procedure described for the adaptive authority experiments is also provided in *SI Appendix*. Mathematica source code verifying the derivations and Python source code for the simulations are openly available at https://doi.org/10.5281/zenodo.15150163.

Data, Materials, and Software Availability. Software data have been deposited in Zenodo (https://doi.org/10.5281/zenodo.15150163) (57). All other data are included in the manuscript and/or *SI Appendix*.

ACKNOWLEDGMENTS. We thank our reviewers for their valuable feedback and pointers to relevant contemporary examples and empirical studies of self-censorship. We also thank Jedidiah Crandall for helpful discussions about Internet censorship technologies, David Peterson for his inputs on the preliminary networked population simulations, Anish Nahar for his preliminary simulation results, and Garrett Parzych for his preliminary analyses of adaptive authorities. J.J.D. is supported in part by the NSF (CCF-2312537). S.F. is supported in part by the NSF (CCF-2211750), ARPA-H (SP4701-23-C-0074), and the Santa Fe Institute. R.A. thanks the University of Michigan for research support.

Downloaded from https://www.pnas.org by Casey Luskin on January 13, 2026 from IP address 72.132.200.2.

1. A. Hsiao, A. Lim, *The Verso Book of Dissent: Revolutionary Words From Three Millennia of Rebellion and Resistance* (Verso, London, United Kingdom, 2020).
2. T. Lokot, *Beyond the Protest Square: Digital Media and Augmented Dissent, Protest, Media and Culture* (Rowman & Littlefield, Lanham, MD, USA, 2021).
3. S. Feldstein, *The Rise of Digital Repression: How Technology Is Reshaping Power, Politics, and Resistance* (Oxford University Press, New York, NY, USA, 2021).
4. P. R. Kalluri *et al*., Computer-vision research powers surveillance technology. *Nature* **643**, 73–79 (2025).
5. P. Mozur, *In Hong Kong Protests, Faces Become Weapons* (The New York Times, 2019).
6. J. Vincent, *NYPD Used Facial Recognition to Track Down Black Lives Matter Activist* (The Verge, 2020).
7. Amnesty International, "Israel and occupied palestinian territories: Automated apartheid: How facial recognition fragments, segregates and controls Palestinians in the OPT" (Tech. Rep. 15/6701/2023, MDE, 2023).
8. A. Loewenstein, *The Palestine Laboratory: How Israel Exports the Technology of Occupation Around the World* (Verso Books, London, United Kingdom, 2023).
9. D. Loucaides, *The Changing Face of Protest* (Rest of World, 2024).
10. C. Morris, *Why Facial Recognition Technology Makes These Campus Protests Different from Those in the Past* (Fast Company, 2024).
11. R. Ensafi, P. Winter, A. Mueen, J. R. Crandall, "Analyzing the great firewall of China over space and time" in *Proceedings on Privacy Enhancing Technologies* (De Gruyter Open, 2015), vol. 2015, pp. 61–76.
12. M. E. Roberts, *Censored: Distraction and Diversion Inside China's Great Firewall* (Princeton University Press, 2018).
13. R. Fedasiuk, Buying silence: The price of internet censorship in China. *China Brief* **21**, 19–25 (2021).
14. A. Epifanova, *Subjugating RuNet: The Kremlin's New Levers of Control over Elections and Society* (The Russia File, 2021).
15. D. Xue *et al*., "TSPU: Russia's decentralized censorship system" in *Proceedings of the 22nd ACM Internet Measurement Conference* (ACM, Nice, France, 2022), pp. 179–194.
16. Cyberspace Administration of China, "This year, the Guangdong Internet Police has dealt with 269,000 pieces of illegal information of all kinds" (2015). https://www.cac.gov.cn/2015-09/28/c_1116702824.htm. Accessed 8 May 2024.
17. T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press, New Haven, CT, USA, 2018).
18. S. Myers West, Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media Soc.* **20**, 4366–4383 (2018).
19. M. F. Common, Fear the reaper: How content moderation rules are enforced on social media. *Int. Rev. Law Comput. Technol.* **34**, 126–152 (2020).
20. R. A. Wilson, M. K. Land, Hate speech on social media: Content moderation in context. *Conn. Law Rev.* **52**, 1029–1076 (2021).
21. F. A. Chowdhury, D. Saha, M. R. Hasan, K. Saha, A. Mueen, *Examining Factors Associated with Twitter Account Suspension Following the 2020 U.S.* (Virtual Event, Netherlands, 2022), pp. 607–612.
22. G. Nicholas, Shedding light on shadowbanning (Center for Democracy & Technology, 2022). https://cdt.org/insights/shedding-light-on-shadowbanning/. Accessed 10 December 2022.
23. C. Toraman, F. Şahinuç, E. H. Yilmaz, "BlackLivesMatter 2020: An analysis of deleted and suspended users in Twitter" in *14th ACM Web Science Conference 2022* (ACM, Barcelona, Spain, 2022), pp. 290–295.
24. Human Rights Watch, Meta's broken promises: Systemic censorship of Palestine content on Instagram and Facebook (2023). https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and. Accessed 9 May 2024.
25. F. Pierri, L. Luceri, E. Chen, E. Ferrara, How does Twitter account moderation work? Dynamics of account creation and suspension on Twitter during major geopolitical events. *EPJ Data Sci.* **12**, 43 (2023).
26. F. Shahid, A. Vashistha, *Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?* (ACM, Hamburg, Germany, 2023), pp. 1–18.
27. J. Kaplan, *More Free Speech and Fewer Mistakes* (2025). https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/. Accessed 13 March 2025.
28. Office of the Director of National Intelligence, Annual Statistical Transparency Report Regarding the Intelligence Community's Use of National Security Surveillance Authorities: Calendar Year 2023 (2024). https://www.dni.gov/index.php/newsroom/reports-publications/reports-publications-2024/3811-annual-statistical-transparency-report-regarding-national-security-authorities-calendar-year-2023. Accessed 9 May 2024.
29. D. J. Trump, *Restoring Freedom Of Speech And Ending Federal Censorship* (The White House, Washington, DC, USA, 2025).
30. D. J. Trump, *Ending Radical And Wasteful Government DEI Programs And Preferencing* (The White House, Washington, DC, USA, 2025).
31. D. J. Trump, *Ending Illegal Discrimination And Restoring Merit-Based Opportunity* (The White House, Washington, DC, USA, 2025).
32. G. C. Loury, Self-censorship in public discourse: A theory of "political correctness" and related phenomena. *Rational. Soc.* **6**, 428–461 (1994).
33. J. Horton, Self-censorship. *Res. Publica* **17**, 91–106 (2011).
34. D. Bar-Tal, Self-censorship as a socio-political-psychological phenomenon: Conception and research. *Polit. Psychol.* **38**, 37–65 (2017).
35. X. Shen, R. Truex, In search of self-censorship. *Br. J. Polit. Sci.* **51**, 1672–1684 (2021).
36. P. Ackerman, B. Rodal, The strategic dimensions of civil resistance. *Survival* **50**, 111–126 (2008).
37. T. C. Schelling, *Some Questions on Civilian Defense in Conflict: Violence and Nonviolence* (Taylor and Francis, London, UK, ed. 1, 1971).
38. G. Sharp, *The Politics of Nonviolent Action* (Porter Sargent, Boston, MA, USA, 1973).
39. E. Chenoweth, M. J. Stephan, *Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict, Columbia Studies in Terrorism and Irregular Warfare* (Columbia University Press, New York, NY, USA, 2011).
40. E. Chenoweth, *Civil Resistance: What Everyone Needs to Know, What Everyone Needs To Know* (Oxford University Press, New York, NY, USA, 2021).
41. YouTube, *YouTube Community Guidelines: Taking Action on Violations* (2025). https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#taking-action-on-violations. Accessed 17 February 2025.
42. J. W. Penney, Chilling effects: Online surveillance and Wikipedia use. *Berkeley Technol. Law J.* **31**, 118–182 (2016).
43. J. W. Penney, Internet surveillance, regulation, and chilling effects online: A comparative case study. *Int. Policy Rev.* **6**, 1–24 (2017).
44. P. Link, "China: The anaconda in the chandelier" in *The New York Review of Books* (2002), vol. 49.
45. Y. R. Zhou, The lying flat movement, global youth, and globality: A case of collective reading on Reddit. *Globalizations* **20**, 679–695 (2023).
46. R. Wintrobe, *The Political Economy of Dictatorship* (Cambridge University Press, Cambridge, MA, USA, 1998).
47. M. Wang, J. Mayer, Self-censorship under law: A case study of the Hong Kong national security law. *Free Open Commun. Internet* **1**, 46–55 (2023).
48. B. Mathew *et al*., Hate begets Hate: A temporal study of hate speech. *Proc. ACM Hum. Comput. Interact.* **4**, 1–24 (2020).
49. M. Mitchell, J. H. Holland, S. Forrest, "When will a genetic algorithm outperform hill climbing?" in *Advances in Neural Information Processing Systems*, J. Cowan, G. Tesauro, J. Alspector, Eds. (Morgan Kaufmann, 1993), vol. 6, pp. 51–58.
50. R. Macfarquhar, T. Cheek, E. Wu, *The Hundred Flowers in The Secret Speeches of Chairman Mao* (Harvard University Asia Center, 1989), pp. 105–111.
51. C. Lin, *The Transformation of Chinese Socialism* (Duke University Press, Durham [N.C.], 2006).
52. K. C. Yang, The aborted Green dam-youth escort censor-ware project in China: A case study of emerging civic participation in China's internet policy-making process. *Telematics Inform.* **28**, 101–111 (2011).
53. A. Cunningham, "Microsoft is reworking Recall after researchers point out its security problems." *Ars Technica* (2024). https://arstechnica.com/gadgets/2024/06/microsoft-makes-recall-feature-off-by-default-after-security-and-privacy-backlash/. Accessed 28 July 2025.
54. A. Cunningham, "Windows Recall demands an extraordinary level of trust that Microsoft hasn't earned." *Ars Technica* (2024). https://arstechnica.com/ai/2024/06/windows-recall-demands-an-extraordinary-level-of-trust-that-microsoft-hasnt-earned/. Accessed 28 July 2025.
55. T. Snyder, *On Tyranny: Twenty Lessons from the Twentieth Century* (Crown, New York, NY, USA, 2017).
56. J. W. Penney, Understanding chilling effects. *Minnesota Law Rev.* **106**, 1451–1530 (2022).
57. J. J. Daymude, A. Nahar, Data from "DaymudeLab/DissentSelfCensorship: v1.0.0." Zenodo. https://doi.org/10.5281/zenodo.15150163. Deposited 4 April 2025.

Downloaded from https://www.pnas.org by Casey Luskin on January 13, 2026 from IP address 72.132.200.2.