Deterring Transgressions at the Apex of Government*

Weijia Li† Yang Xie‡

June 18, 2025

Abstract

We analyze a model of deterring transgressions by the head of government, in which the conditions determining whether a transgression will be punished only become clear after it occurs. We identify a key challenge in sustaining such deterrence: effective deterrence prevents new information about the underlying conditions from being revealed, over time rendering the threat of punishment not credible enough to deter transgressions that are sufficiently profitable for the leader. We demonstrate that several institutional or cultural solutions either fail to overcome this challenge or face inherent limitations. The same mechanism also applies to deterring other undesirable behaviors when the precise conditions for punishment only become clear after the fact, such as international aggression, and to maintaining other preventive policies whose success could undermine their perceived necessity, such as anti-discrimination measures and vaccination efforts.

*We thank Yunus Aybas, Scott Ashworth, Cuimin Ba, Matheus Bandeira, Chris Bidner, Alberto Bisin, Patrick Bolton, Kyle Chauvin, Liuchun Deng, Gabriele Gratton, Guojun He, Da Huang, Aziz Huq, Alexander Jakobsen, Ye Jin, Matthew Kahn, Ethan Kaplan, Urmee Khan, Zhao Liu, Zhaotian Luo, Aprajit Mahajan, Roger Myerson, Mariann Ollár, Marcin Pęski, Ariel Roginsky, Gérard Roland, Konstantin Sonin, Yufeng Sun, Michael Ting, Scott Tyson, Fan Wang, Shaoda Wang, Melanie Wasserman, Yinxi Xie, Siyang Xiong, Dingxin Zhao, Congyi Zhou, and participants in seminars at Monash, NYU Shanghai, Seikei, and THU and 2023 SIOE and 2024 GAMES Meetings for their valuable comments. Earlier versions of the paper were titled "Forever Deterred, the Highest Crimes? A Difficulty and Potential Solutions" or "Can High Crimes Be Perpetually Deterred? A Difficulty and Potential Solutions."

†Department of Economics, Monash University; weijia.li@monash.edu.

‡Department of Economics, University of California, Riverside; yang.xie@ucr.edu.

1 Introduction

Transgressions by leaders at the apex of government against legal or political norms can impose "truly massive costs" on society (Huq, 2018, p. 1510; Rivera et al., 2025). Deterring such acts often relies on political accountability mechanisms, such as elections, impeachment, or backlash from key supporters (e.g., Madison, 1845, p. 528; Myerson, 2008, 2019; Ginsburg et al., 2021). This presents a challenge: if these political threats of punishment are credible, transgressions are deterred – so the threats are never carried out. But if they are never carried out, how can they remain credible over time?

In this paper, we formalize and analyze this challenge in a model of presidential transgression and impeachment. In the model, a President (P) in each period chooses whether to transgress, and a Congress (C) decides whether to punish the transgressing P through impeachment. Main insights from the model also apply to deterring transgressions by top leaders in non-presidential or non-democratic regimes, as P can represent the leader in question – a prime minister, ruling party general secretary, or monarch; impeachment thus stands for the relevant political accountability mechanism, which C implements.

The key feature of our model is that the incentive for C to impeach the transgressing P – referred to as the state of the world – follows a stochastic process, and the current state of the world is only known to the players after P transgresses. This reflects the reality that, given the political nature of impeachment, or more broadly, punishment for transgressions by the head of government, it is always difficult to fully predict ex ante whether or not such punishment will be carried out. This feature implies that successful deterrence prevents new empirical knowledge about the state of the world from emerging. As a result, over time, in a generally uncertain yet path-dependent world, this lack of new information renders historical knowledge from the times of past transgressions increasingly irrelevant to the current state of the world. Consequently, the credibility of the threat of impeachment becomes increasingly akin to a coin toss with a fixed probability. Therefore, as long as transgression is sufficiently profitable for P, there will always come a time when the threat becomes not credible enough to deter transgression. Deterrence is thus dynamically self-undermining, and this formalizes the challenge in deterring transgressions at the apex of government.

Further analysis reveals that this challenge arises distinctively from our model: it is transgressions by the head of government, not misconduct or crimes in ordinary settings, that are in question; the state of the world evolves in an uncertain yet path-dependent manner; and the uncertainty centers on whether C will impeach a transgressing P, rather than P's payoffs from transgressing. We also show that this challenge persists even if part of the state of the world is perfectly predictable ex ante, provided that it is bounded, and

the challenge applies more strongly when the transgression in question is truly exceptional and fundamentally distinct from those minor ones.

We then examine several potential institutional or cultural solutions within our framework, finding that each of them faces limitations. First, while equipping impeachment with a sufficiently severe punishment can make perpetual deterrence possible, the required severity may incentivize P to avoid punishment, to the extent that it may encourage the transgressing P to execute a self-coup. Second, admitting many non-partisan members to Congress with a super-majority rule for conviction, which is common in practice, may exacerbate the problem, making perpetual deterrence impossible even with an arbitrarily harsh punishment. Third, establishing an additional punitive authority, such as an independent judiciary, can only partially alleviate the problem. Fourth, subjecting the transgressions in question to many overlapping punitive authorities, as in federalism, could be an answer to the challenge, but this requires all these authorities to be independent from each other – a condition that may be unrealistic – and risks creating political conflict due to inconsistent decisions or procedures. Finally, fostering a political culture against transgressions by the head of government cannot fully address the challenge, as long as the influence of such efforts decays, however slightly, over time.

Since Becker (1968), the economics literature on deterrence of crimes and other illegal behaviors has primarily focused on settings that are more ordinary than the apex of government, showing theoretically and empirically that crimes in such settings are more responsive to the certainty of apprehension than to the severity of punishment (e.g., survey by Chalfin and McCrary, 2017).[1] Our paper suggests that the logic of deterring transgressions at the apex of government can be fundamentally different. This is because, other than the highest office of government, few alternative settings exist where the underlying circumstances of such transgressions can be learned. As a result, in the long run, maintaining the certainty of apprehension for transgressions by the head of government can be infeasible.

About such transgressions, a thread of literature in political economy has focused on institutions that deter abuse of power through credible threats of punishment (e.g., Przeworski, 2006; Weingast, 1997; Myerson, 2008, 2019; Fearon, 2011; Svolik, 2012; Gitmez et al., 2023). We show that deterrents based on such threats may be undermined by their own success over time. This challenge is robust to several proposed institutional solutions, and to the cultural approach advocated in the constitutional law literature (e.g., Huq, 2018).

Recent history shows that even long-established, advanced democracies are not immune to transgressions by their leaders against democratic institutions and norms (e.g., Bow-

[1] A long tradition in law and criminology also maintains that crimes in ordinary settings are more responsive to the certainty of punishment than to its severity (e.g., Beccaria, 2008, p. 49; Nagin, 2013).

man, 2024). This challenges the conventional wisdom that prolonged democratic governance fosters immunity to such transgressions (e.g., Przeworski, 2006; Fearon, 2011; Bidner and Francois, 2013; Besley and Persson, 2019; Acemoğlu and Robinson, 2022; Acemoğlu et al., 2025). We reconcile this observation with the literature by showing that constant deterrence of such transgressions can become its own enemy, a possibility rooted in their unique setting, i.e., the apex of government, and the political nature of their punishment.

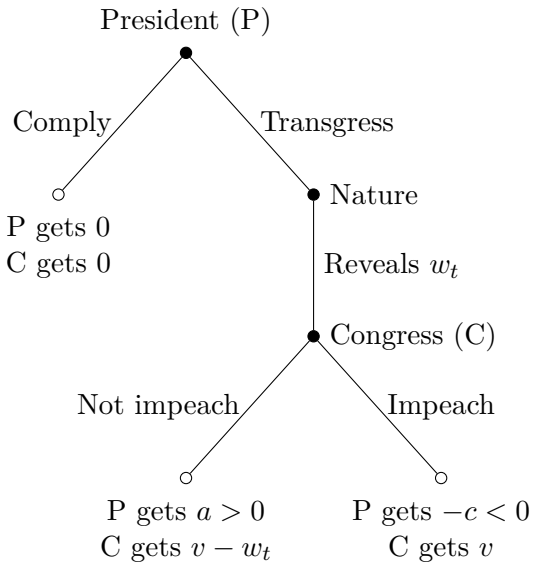
Interpreting transgressions in our model as attacks on democratic norms and executive constraints links our paper to the growing literature on democratic backsliding and executive transgressions (e.g., Acemoğlu et al., 2013; Howell and Wolton, 2018; Svolik, 2020; Chiopris et al., 2021; Matějka and Tabellini, 2021; Miller, 2021; Grillo and Prato, 2023; Luo and Przeworski, 2023; Gratton and Lee, 2024; survey by Grillo et al., 2024). Our model generates stochastic cycles between periods of constant deterrence and episodes of consecutive transgressions, i.e., backsliding, while each transgression becomes a critical juncture, revealing for how long the threat of punishment can remain credible to prevent backsliding. Unlike most studies in this literature, our model does not feature bounded rationality or informational disadvantages on the side of the custodian of democracy, i.e., C in our model; neither do the dynamics generated rely on such features.

More broadly, our paper contributes to the economics literature on the consequences of knowledge deficiency, which has focused on the implications of *exogenous* lacks of knowledge for managing central planning, catastrophic risks, and power relations (e.g., von Hayek, 1945; Weitzman, 1974, 2009, 2011; Stern, 2013; Huang et al., 2017; Conte and Kelly, 2021; Li et al., 2022). In our setting, knowledge about the state of the world is *endogenously* hidden by successful deterrence, a desirable outcome. A prolonged period of this desirable outcome may eventually lead to its opposite: transgression. This mechanism also applies to deterring other undesirable behaviors when the precise conditions for punishment only become clear after the fact, and to other contexts where the success of a preventive policy may hinder updated understanding of its necessity, thus threatening its continuation. We discuss three examples of such applications at the end of the paper: deterring international aggression, upholding or not the United States Voting Rights Act of 1965, and vaccine hesitancy.

The paper proceeds as follows. Section 2 presents the baseline model. Section 3 analyzes it and delivers the main result. Section 4 examines proposed institutional and cultural solutions. Section 5 concludes by discussing applications of our model to other contexts. Proofs, technical details, and extensions are provided in the Appendix.

2 Baseline Model

In each period t, a President (P) and a Congress (C) interact in a game as in Figure 1. P first decides whether to comply with or transgress the law and norms governing the presidency. If P complies, both P and C receive a payoff of zero, and the period ends, transitioning to period $t + 1$.



Players know time, $s < t$, and state of the world, w_s, at last transgression.

Figure 1: Baseline model, period t

If P transgresses, C will subsequently decide whether to punish P through impeachment. Modeling impeachment as punishment aligns with the constitutional law literature, which views punishment as a central purpose of impeachment (e.g., Tribe and Matz, 2018; Crespo, 2019). If C impeaches P, P incurs an exogenous cost, $c > 0$, while C receives an exogenous payoff, v. If C chooses not to impeach P, P receives an exogenous, positive gain, $a > 0$, while C receives $v - w_t$, where w_t represents the net incentive for C to impeach P. The period then ends, transitioning to period $t + 1$.

State of the world. We interpret w_t as the state of the world at period t. This approach aligns with insights from comparative and constitutional law. Impeachment, or more broadly, punishment for transgressions by the head of government, is ultimately a political solution to political crises where the transgressions in question have significantly undermined public confidence in the incumbent or threatened constitutional governance. Whether such punishment will be carried out is thus primarily a political question, and its answer depends on many specific circumstances at the time, be them political, social, economic, or

moral. This holds true even when supposedly apolitical experts, such as prosecutors and judges, are involved (e.g., Baumgartner and Kada, 2003; Tribe and Matz, 2018; Crespo, 2019; Ohnesorge, 2020; Ginsburg et al., 2021; Bowman, 2024; Monaghan et al., 2024).

Conditional revelation. The dependence on these specific circumstances suggests that, while many factors, such as the alignment of electoral interests between P and C, may be known or predictable before a transgression occurs, there is always a component of these circumstances that is difficult to fully predict ex ante. Additionally, constitutional law is inherently general and often leaves punishable transgressions by the head of government incompletely defined, making it difficult to predict beforehand how a transgression will be addressed (e.g., Story, 1833, p. 287; Huq, 2018, p. 1515; Tribe and Matz, 2018, p. 45; Bowman, 2024, p. 99).[2] As a result, one today has to estimate whether a transgression will be punished, while relying on her understanding of the circumstances of past transgressions (e.g., Tribe and Matz, 2018; Congressional Research Service, 2023, p. 839; Bowman, 2024).

This leads to the key feature of our model: the conditional revelation of the state of the world. As in Figure 1, at the start of period t, the players do not know the exact value of w_t; once P chooses to transgress, nature reveals w_t to them. In Section 3.2, we discuss the robustness of our analysis to including an additional known component of the state of the world at the start of each period.

Apex of government vs. ordinary settings. Such conditional revelation distinguishes transgressions at the apex of government from misconduct and crimes in ordinary settings. For such misconduct or crimes, players typically have abundant ex ante knowledge about the state of the world, as there are many alternative settings to the setting in question to learn from. In Section 3.2, we explore the dynamics of transgression and punishment when the transgression occurred in an ordinary setting, not at the apex of government.

Evolution of the state of the world. At the start of period t, the players know when past transgressions occurred and the corresponding states of the world. How they use this

[2] A notable example is Section 4 of Article II of the Constitution of the United States, which states: "[t]he President, Vice President and all civil Officers of the United States, shall be removed from Office on Impeachment for, and Conviction of, Treason, Bribery, or other high Crimes and Misdemeanors." The term "high Crimes and Misdemeanors" is a "catchall phrase" and "not defined in the Constitution or statutes," creating "considerable vagueness" (Black and Bobbitt, 2018, p. 26; Congressional Research Service, 2023, p. 839). For further literature, see Tribe (1998), Tribe and Matz (2018), and Bowman (2024). For theoretical arguments for constitutional law to remain silent on this issue, see Huq (2018) and Crespo (2019).

information depends on how w_t evolves. We assume that w_t follows a Markov process,

$$w_t = \rho w_{t-1} + \epsilon_t, \text{ where } 0 < \rho \leq 1, \text{ and } \epsilon_t \sim \mathcal{N}(0, \sigma^2), \text{ i.i.d., with } \sigma > 0. \tag{1}$$

Here, ϵ_ts represent contemporary shocks, which are mutually independent and identically and normally distributed, with a mean of zero. This reflects our approach that w_t captures the component of the state of the world that is unknown and difficult to predict ex ante. We assume $\sigma > 0$ and $0 < \rho \leq 1$ because the state of the world relevant to the transgressions in question is most reasonably uncertain yet path-dependent, but not oscillatory or explosive. In particular, we allow the influence of historical shocks to either decay ($0 < \rho < 1$) or persist ($\rho = 1$) over time. In Sections 3.2 and 4.5, we discuss alternative assumptions about the evolution of w_t, such as alternative ranges of ρ or σ, or including a drift term.

Complete the setup. For simplicity, we assume that players are myopic, caring only about their payoffs within the current period. This is justifiable if each period represents the finite term of office for the players. We discuss the robustness of our results to forward-looking players in Section 3.2.

We solve the model using backward induction. For simplicity, we assume that C will not impeach a transgressing P if indifferent between impeaching and not impeaching, and P will comply if indifferent between transgressing and complying.

3 Analysis and Main Result

We now solve the game at period t. C will impeach a transgressing P if and only if $w_t > 0$. Since w_t follows a Markov process, the time of the last transgression, $s < t$, and the state of the world at that time, w_s, provide all the information P needs to estimate w_t at the beginning of period t. The estimated probability of punishment is $\mathbf{P}[w_t > 0 \mid w_s, s]$. Thus, P transgresses if and only if

$$a \cdot (1 - \mathbf{P}\left[w_t > 0 \mid w_s, s\right]) - c \cdot \mathbf{P}[w_t > 0 \mid w_s, s] > 0, \tag{2}$$

or equivalently, if the conditional probability of punishment is sufficiently low:

$$\mathbf{P}[w_t > 0 \mid w_s, s] < a/(a + c). \tag{3}$$

We first examine how the conditional probability of punishment, $\mathbf{P}[w_t > 0 \mid w_s, s]$, depends on both the time elapsed since the last transgression, $t - s$, and the state of the

world revealed at that time, w_s. The evolution of w_t implies that

$$w_t = \rho^{t-s} w_s + \sum_{\tau=s+1}^{t} \rho^{t-\tau} \epsilon_\tau. \tag{4}$$

The conditional distribution of w_t is thus

$$(w_t \mid w_s, s) \sim \begin{cases} \mathcal{N}\left(\rho^{t-s} w_s, \left(1-\rho^{2(t-s)}\right) \cdot \sigma^2 / \left(1-\rho^2\right)\right) & \text{if } \rho \in (0,1); \\ \mathcal{N}\left(w_s, (t-s) \cdot \sigma^2\right) & \text{if } \rho = 1. \end{cases} \tag{5}$$

This implies that the conditional probability of punishment is

$$\mathbf{P}[w_t > 0 \mid w_s, s] = \begin{cases} \Phi\left(\rho^{t-s} w_s \Big/ \left(\sigma \cdot \sqrt{\left(1-\rho^{2(t-s)}\right) / \left(1-\rho^2\right)}\right)\right) & \text{if } \rho \in (0,1); \\ \Phi\left(w_s \big/ \left(\sigma \cdot \sqrt{t-s}\right)\right) & \text{if } \rho = 1, \end{cases} \tag{6}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. This conditional probability has the following properties:

Lemma 1 (Credibility of impeachment threat)**.** *It is more likely than not for C to impeach a transgressing P if and only if C impeached P at the last transgression. As time passes with constant deterrence, the threat to punish increasingly resembles a 50/50 coin toss. Mathematically, for any $t > s$,*

- *if $w_s \leq 0$, then $\mathbf{P}[w_t > 0 \mid w_s, s] \leq 0.5$, increases with $t - s$, and converges to 0.5 as $t - s$ increases;*
- *if $w_s > 0$, then $\mathbf{P}[w_t > 0 \mid w_s, s] > 0.5$, decreases with $t - s$, and converges to 0.5 as $t - s$ increases.*

The intuition behind Lemma 1 is as follows. First, why does whether C impeached P at the last transgression, i.e., whether $w_s > 0$, indicate whether it is more likely than not for C to impeach a transgressing P now? This is because, since the world is path-dependent and no new empirical knowledge about the state of the world has been revealed since the last transgression, the current estimate of the state of the world must be, adjusted with the degree of path-dependence, centered on the last revealed state.

Second, why is the monotonic convergence to a 50/50 coin toss? In the case of $\rho \in (0,1)$, this is because the influence of historical shocks decays over time, which is set mechanically by $\rho \in (0,1)$. As time passes with constant deterrence, because of the conditional revelation of the state of the world, no new information about the state of the world has been revealed.

The conditional distribution of the current state of the world thus converges to a normal distribution with a mean of zero and a finite variance, i.e.,

$$(w_t \mid w_s, s) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2/\left(1-\rho^2\right)\right) \text{ as } t-s \to \infty, \quad \text{if } \rho \in (0,1). \tag{7}$$

This makes the threat of impeachment increasingly resemble a 50/50 coin toss.

In the case of $\rho = 1$, even though the influence of historical shocks would remain constant, its relevance would dwindle, still. This is because, without new revelation, the uncertainty in the evolution of the state of the world piles up over time and is not bounded from above. The conditional distribution of the state of the world thus converges to a normal distribution with an infinite variance, i.e.,

$$(w_t \mid w_s, s) \xrightarrow{d} \mathcal{N}\left(w_s, \infty\right) \text{ as } t-s \to \infty, \quad \text{if } \rho = 1. \tag{8}$$

This makes the threat of impeachment increasingly resemble a 50/50 coin toss, too.

Lemma 1 implies that, to understand the dynamics of impeachment and transgression under constant deterrence, we need to compare the critical threshold, $a/(a+c)$, with the conditional probability in the limit over time with constant deterrence, which is one-half. On that, $a/(a+c)$ has the following property:

Lemma 2 (50/50 coin toss)**.** *A threat of impeachment equivalent to a 50/50 coin toss cannot deter P from transgressing, if and only if the potential gain from transgression exceeds the punishment from impeachment. Mathematically,*

$$0.5 < a/(a+c) \text{ if and only if } a > c. \tag{9}$$

The intuition behind Lemma 2 is straightforward. If P perceives the threat of impeachment as a 50/50 coin toss, he will simply compare the gain from transgression with the punishment, and will transgress if and only if the gain is greater, however slightly.

Combining Lemmas 1 and 2, we derive the main result of this paper:

Proposition 1 (Challenge in deterring transgressions at the apex of government)**.** *For any $a > c$, define $\bar{w} \equiv (\sigma/\rho) \cdot \Phi^{-1}\left(a/(a+c)\right) > 0$. Then,*

1. *if $w_s \leq 0$, then C will not impeach P at period s, and P will transgress at $s+1$;*
2. *if $0 < w_s < \bar{w}$, then C will impeach P at s, but P will still transgress at $s+1$;*
3. *if $w_s \geq \bar{w}$, then C will impeach P at s, and there exists a period $T \in [s+2, \infty)$, such that P will comply from $s+1$ to $T-1$ but will transgress at T.*

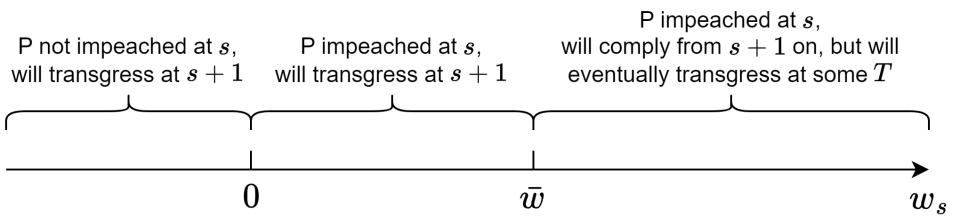


Figure 2: Proposition 1: Perpetual deterrence impossible for any $a > c$

Figure 2 visualizes Proposition 1. We prove Proposition 1 in Appendix A. The intuition is as follows. About Claim 1, by Lemma 1, an unpunished transgression ($w_s \leq 0$) suggests that the impeachment threat in the following period is no more credible than a 50/50 coin toss. By Lemma 2, this is not enough to deter another transgression, whenever the potential gain from transgression exceeds the punishment ($a > c$).

About Claim 2, having the last transgression punished ($w_s > 0$) may not be enough to deter transgression in the following period, since, by Lemma 1, the credibility of the threat of impeachment may drop too much even in just one period. Therefore, the last revealed state of the world has to be sufficiently against transgression ($w_s \geq \bar{w}$) to deter another transgression right after the last transgression.

About Claim 3, by Lemma 1, even if the threat of impeachment deters transgression right after the last transgression ($w_s \geq \bar{w}$), as time passes with constant deterrence, the threat increasingly resembles a 50/50 coin toss. By Lemma 2, a threat equivalent to a 50/50 coin toss cannot deter transgression, as long as the potential gain from transgression exceeds the punishment. Therefore, there must exist a period when the threat will eventually become not credible enough to deter transgression.

The challenge, formalized. Proposition 1 formalizes the challenge in deterring transgressions at the apex of government: for any transgression that is sufficiently profitable, constant deterrence would over time make the threat of punishment as non-credible as a coin toss with a fixed probability, i.e., 50/50 in our baseline model, eventually leading to a failure of deterrence.

3.1 Numerical Illustration

To illustrate the dynamics implied by Proposition 1, we provide a numerical simulation. Figure 3 plots the typical dynamics over 500 periods. In this simulation, the initial state of the world w_1 is revealed to everyone before period 2 and is set exogenously to be sufficiently against transgression, i.e., $w_1 \geq \bar{w} > 0$, ensuring that P does not transgress at period 2.

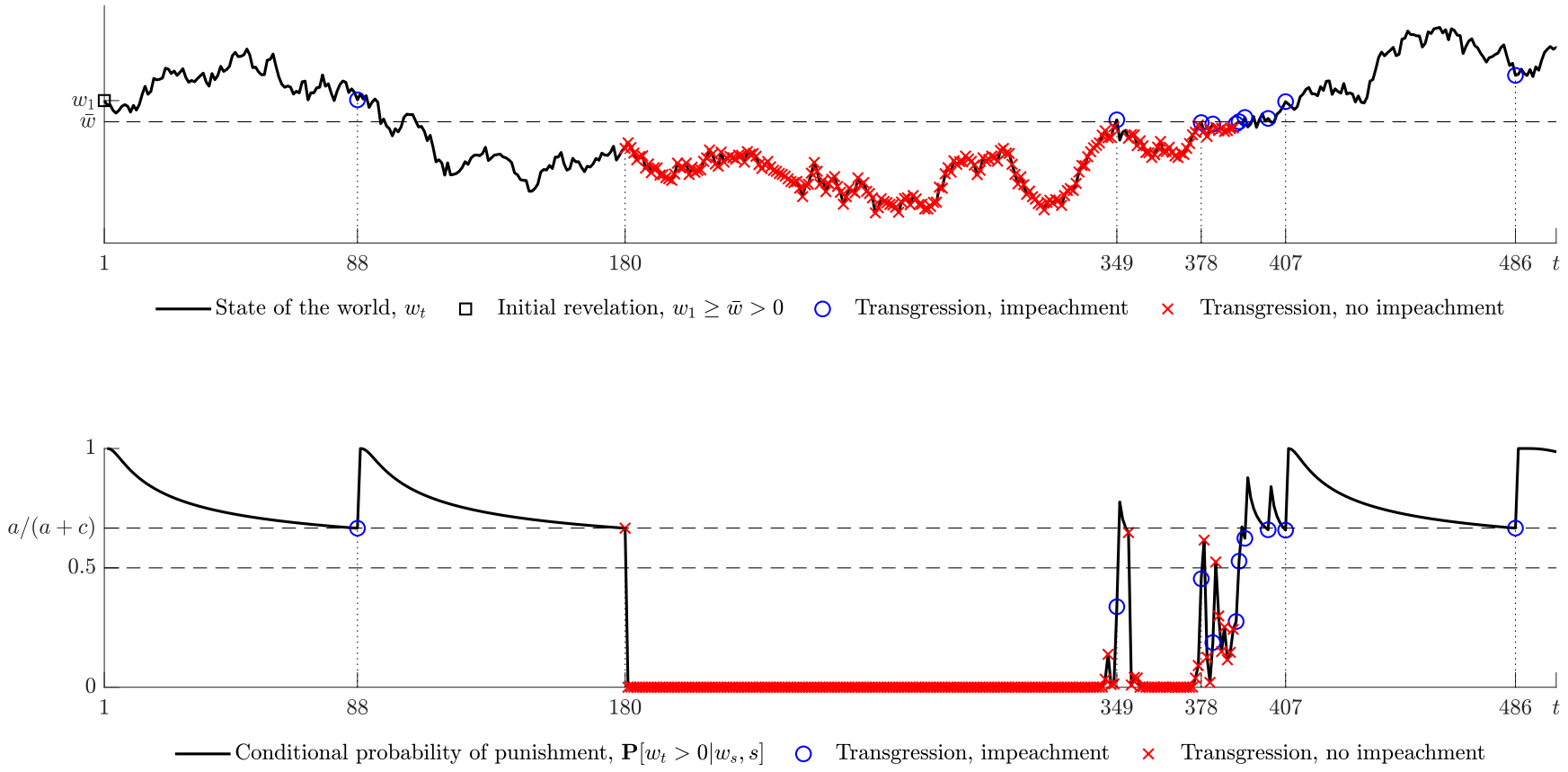


Figure 3: Typical dynamics with transgression gain exceeding punishment ($a > c$)

Consistent with Lemma 1, the conditional probability of punishment declines over time with constant deterrence. Consistent with Proposition 1, at period 88, the threat of impeachment becomes not credible enough to deter transgression, i.e., $\mathbf{P}[w_{88} \mid w_1, 1] < a/(a+c)$. At this point, the revealed state of the world is strongly against transgression, i.e., $w_{88} \geq \bar{w}$, leading to another lengthy period of deterrence. Yet, consistent with Lemma 1, the conditional probability of punishment still declines over time due to constant deterrence.

At period 180, the threat of punishment again becomes not credible enough to deter transgression. This time, the revealed state of the world is too low even for C to impeach P, i.e., $w_{180} \leq 0$. Consistent with Proposition 1, this leads to another transgression at period 181. The revealed state of the world again prevents C from impeaching P. This pattern of consecutive transgressions without being punished continues until period 349.

From period 349 onward, transgressions still occur frequently. There are periods, such as period 378, when the transgressing P is impeached, but the revealed state of the world cannot sustain deterrence. It is only at period 407 that the world has moved sufficiently against transgression, leading to a lengthy period of deterrence until period 486. The revelation at period 486 is even more strongly against transgression, starting a longer period of deterrence. That said, as Proposition 1 predicts, another transgression will eventually occur.

Stochastic cycles and critical junctures. This illustration suggests that stochastic cycles between constant deterrence and consecutive transgressions may emerge. Each transgression can be viewed as a critical juncture, indicative of future dynamics. If the revealed state of the world is not strongly against transgression – particularly if it prevents C from impeaching P – it may lead to a prolonged period of consecutive transgressions. Conversely, if the revelation is strongly against transgression, it will secure punishment for the current transgression and deter potential transgressions for many subsequent periods. Indeed, we show in Appendix A that the length of each period of constant deterrence, T in Proposition 1, increases with w_s.

3.2 Discussion

To better understand the main result, we discuss several alternative setups of the model:

Not at the apex of government. What if the transgression in question is not committed by the head of government but in an ordinary setting? As discussed in Section 2, this would correspond to a setup where the players always know the value of w_t at the start of period t, which we formalize in Appendix B. In this case, constant deterrence would not imply a lack of new empirical knowledge about the state of the world. The potential transgressor's

decision would only depend on w_t; once conditional on w_t, past transgression and punishment decisions would have no effect on the current transgression decision. Thus, deterrence would not be self-undermining over time, and the challenge in Proposition 1 would not exist.

An additional known component. What if part of the state of the world is known or perfectly predictable ex ante? In Appendix C, we introduce an additional component to the net incentive for C to impeach P that is independent of w_t and publicly known at the beginning of each period. We show that, as long as this component is bounded, i.e., if social knowledge about the state of the world is never too wild ex ante, over time with constant deterrence, the credibility of the threat of impeachment will still converge to a probability strictly lower than one. Therefore, the challenge in Proposition 1 will persist, though the required threshold for the gain from transgression, i.e., $a > c$ in Proposition 1, may change.

Testing the state of the world by minor transgressions. What if P could test the state of the world by committing a minor transgression in each period before deciding whether to proceed with the "real" transgression? We formalize this idea as an extension of the baseline model in Appendix D. In this extension, in addition to the baseline setup, P commits a costless, minor "test" transgression at the start of each period, which generates an additional public signal about the state of the world. We assume that the precision of this signal depends on how different the real and test transgressions are in nature. Both players then process these signals using Bayesian updating.

We show that, at one extreme, if the real transgression is as minor as the test transgression, the additional signal in each period would perfectly reveal the net incentive for C to impeach P if the real transgression were committed. In this case, the game's dynamics converge to those of deterring a misconduct or crime in an ordinary setting, as discussed above. At the other extreme, if the real transgression is egregious to an unthinkable degree, the signal from the minor test transgression would reveal very little about the current state of the world. Here, the game's dynamics converge to those of our baseline model. Therefore, the challenge in Proposition 1 applies more strongly to scenarios where the transgression in question is truly exceptional and fundamentally distinct from those minor ones.

Alternative ways of evolution of the state of the world. We have assumed that the state of the world is uncertain yet path-dependent, but not oscillatory or explosive, i.e., $\sigma > 0$ and $0 < \rho \leq 1$. What if it evolves differently? We discuss these alternatives below, with technical details provided in Appendix E.

Deterministic world, i.e., $\sigma = 0$. In this case, although constant deterrence would still not bring new empirical knowledge about the state of the world, the revelation at the last transgression would perfectly indicate the current state. Once a transgression is punished, P would understand that any future transgression will also be punished for sure, leading to perpetual deterrence. Thus, the challenge in Proposition 1 would not exist.

Path-independent world, i.e., $\rho = 0$. In this case, although constant deterrence would still not bring new empirical knowledge about the state of the world, the lack of new empirical knowledge would not become more consequential over time, and P's decision to transgress would be independent of history. Thus, deterrence would not be self-undermining over time, and the challenge in Proposition 1 would not exist, either.

Oscillatory world, i.e., $\rho < 0$. In an oscillatory world, absent new empirical knowledge about the state of the world, a threat of impeachment that is more credible than a 50/50 coin toss in one period will become less credible than a 50/50 coin toss in the following period. Therefore, if the gain from transgression exceeds the punishment ($a > c$), deterrence would never last for more than one period. Thus, the challenge in Proposition 1 would have no opportunity to emerge.

Explosive world, i.e., $\rho > 1$. In this case, as time passes with constant deterrence, the expectation and standard deviation of the state of the world, conditional on its value and time at the last transgression, explode at about the same rate, ρ. Therefore, once the state of the world is revealed to be sufficiently against transgression, deterrence can last forever. Thus, the challenge in Proposition 1 would not arise.

Payoff uncertainty. What if the conditional revelation were about P's payoffs from transgressing, rather than C's strategy to punish? In Appendix F, we examine this alternative setup. We assume that C impeaches P with an exogenous probability, while P's gain and punishment for transgression evolve (with the same degree of path-dependence) and are subject to conditional revelation, similar to w_t in the baseline model.

We show that perpetual deterrence is possible in this case. Although the payoff uncertainty accumulates over time with constant deterrence, it does not affect the expectation of the relative gain from transgression to punishment. Therefore, it will not affect whether P will comply or transgress. As a result, deterrence can last forever. The challenge in Proposition 1 thus arises from strategic uncertainty, not payoff uncertainty.

Forward-looking players. What if the players in our model cared not only about their current payoffs, but also their future payoffs? In Appendix G, we analyze a setting in which any impeached P will be replaced by a new P, while C and all Ps are infinitely forward-looking. In the analysis, we focus on symmetric, pure-strategy Markov-perfect equilibria, where "symmetric" means that all Ps adopt the same strategy.

We show that the challenge in Proposition 1 still emerges whenever the one-time gain from transgression exceeds the net present value of the punishment from impeachment. This is because, first, about C's strategy, the state of the world evolves exogenously, and the current state is determined and revealed before C decides whether to impeach a transgressing P. Thus, C's decision about impeachment does not affect the future state of the world, so a forward-looking C behaves just like a myopic one in our baseline model.

Given that, for each forward-looking P, the intuition behind Proposition 1 still applies. Not only that, in equilibrium, P's continuation value following an unpunished transgression must be weakly positive, as P can always choose to unconditionally comply forever, receiving a zero continuation value. This positive continuation value creates an additional incentive for P to transgress now, making constant deterrence even harder to sustain. Therefore, the challenge in Proposition 1 is robust to having forward-looking players in our analysis.

Summary. These discussions clarify that the challenge presented in Proposition 1 arises from 1) a lack of new empirical knowledge about the world due to constant deterrence, and the world being 2) uncertain and 3) path-dependent. Having significant transgressions at the apex of government in question that are fundamentally different from those minor ones implies a lack of new empirical knowledge under constant deterrence; uncertainty makes this lack of knowledge consequential, and path-dependence amplifies its impact over time, eventually making the threat of punishment not credible enough to deter transgression. This self-undermining nature of deterrence is rooted in strategic uncertainty, not payoff uncertainty, and remains robust even when the state of the world includes a known, bounded component, and when the players are forward-looking.

4 Potential Institutional or Cultural Solutions

Having formalized and analyzed the challenge in deterring transgressions by the head of government, we now explore potential institutional and cultural solutions. As we will see, each of them faces limitations.

4.1 Severe Punishment

Proposition 1 addresses the case where P's gain from transgression exceeds the punishment ($a > c$). Could a sufficiently severe punishment solve the problem? In Appendix H, we show that perpetual deterrence is possible if the punishment outweighs the gain from transgression ($c \geq a$). This is because, even though the threat of impeachment converges to a 50/50 coin toss over time with constant deterrence, it would still be sufficient for deterrence if the punishment were severe enough.

But relying on severe punishment as a solution has limitations. First, a classic idea in law and criminology argues that, as put by Beccaria (2008, p. 51), "impunity itself arises from the atrocity of punishments." The more severe the punishment, the harder the offender will try to avoid it, and the more reluctant the punitive authority may become when considering whether to implement the punishment. Consequently, the punishment will be applied less consistently (e.g., Beccaria, 2008, p. 49–57). Not only that, to preempt the punishment, a transgressing P may even choose to execute a self-coup, committing another transgression (e.g., Cameron, 1998, p. 134). Thus, independent of Proposition 1, having a severe punishment can make the threat of punishment less credible, or even encourage further transgressions.

Second, the gain from transgressions by leaders at the apex of government is often enormous in political value, especially when the objective is to control the highest office of government, which wields the power of legitimate violence or proclaiming states of exception (e.g., Weber, 2004; Schmitt, 1985; Agamben, 2005). For the punishment to outweigh such gains, it may need to be extremely harsh, making it infeasible in a civilized society (e.g., Beccaria, 2008, p. 85; Durkheim, 1973).[3]

A third limitation is that, when deterrence has been successful in the past, it may be difficult for outsiders, such as social scientists, journalists, or concerned citizens, to determine whether the punishment is indeed sufficiently severe. Our model helps to elaborate on this point. As shown in Appendix H, an ongoing period of constant deterrence can occur in both cases ($a > c$ and $a \leq c$). Thus, observing an ongoing period of constant deterrence alone cannot reveal whether the punishment is sufficiently harsh to ensure perpetual deterrence.

[3]As an famous example, Priscus Attalus, a leading Roman senator, was maimed of a thumb and a forefinger in 416 for collaborating with the Visigoths and usurping the Roman emperorship in 409 and 414 (e.g., Bury, 1889, p. 150). Such a punishment, combined with public humiliation, "was intended as a lesson for other Roman aristocrats of the dangers of independent action" (Salzman, 2021, p. 108).

4.2 Wisdom of the Crowd

In our baseline model, C is a single player. As "the many are better judges …when they come together" (Aristotle, 1998, p. 83), could the wisdom of the crowd, i.e., letting C consist of many non-partisan members, overcome the challenge in deterring transgressions at the apex of government? In Appendix I, we analyze this setting by incorporating multiple members into Congress, each with their own net incentive to convict a transgressing P, evolving and subject to conditional revelation as w_t in the baseline model. To ensure that these members are non-partisan, we assume that these net incentives are mutually independent. We also assume that a transgressing P will be punished if and only if he is convicted by a vote in Congress, given a voting rule for conviction, and that each member votes sincerely.

We show that, given any super-majority rule for conviction, which is often required by the seriousness of the matter in question, having many non-partisan members in Congress exacerbates the problem presented in Proposition 1. This is because, as in the baseline model, over time with constant deterrence, each Congress member will vote to convict a transgressing P almost like a 50/50 coin toss. Since these members are non-partisan, the Law of Large Numbers applies. Thus, in the limit over time, as the size of Congress increases, it becomes almost certain that only about half of the members will vote to convict, failing to clear the super-majority threshold.[4] Therefore, even if the punishment exceeds the gain from transgression, this almost entirely non-credible threat of punishment will fail to deter P from transgressing.[5] In this sense, for a wide range of reasonable voting rules for conviction, such as any super-majority rule, the wisdom of the crowd does not help overcome the challenge in deterring transgressions at the apex of government.[6]

[4]For example, in this limit, the probability of having at least two thirds of a Congress of 50 non-partisan members voting to convict a transgressing P is 0.0077; the probability for a 100-member Congress is 0.0004. One can also read this result as applying a special case of Condorcet's jury theorem (Condorcet, 1785; Grofman et al., 1983, p. 264), in which each voter votes for the correct decision with probability 0.5, to super-majority voting rules.

[5]Bolton and Rosenthal (2002) recognize that a democratic process (majority or super-majority voting) can certify occasions (widespread financial distress) about which economic contracts (debts) are incomplete and warrant intervention (moratoria), and they show that such an ex post remedy, when anticipated, may improve efficiency of resource allocation ex ante. Along this line, we recognize that, on occasions (transgressions by the head of government) about which social contracts (constitutions) can be incomplete, a democratic process (super-majority voting) can be used to certify a proper intervention (impeachment or not), and we show that such an ex post remedy, when anticipated, will make such occasions inevitable. This provides an interpretation in the framework of incomplete contracts for this result.

[6]One may entertain the idea of having a voting rule that requires a super-majority for acquittal, i.e., a minority rule for conviction. Indeed, we show in Appendix I that, in the limit over time with constant deterrence, under such minority rule, as the size of Congress increases, it becomes almost certain that any transgressing P will be convicted, thus deterring transgression. Yet, in reality, such voting rules may encourage wrongful conviction of complying Ps.

4.3 Additional Independent Punitive Authority

In our baseline model, C is the only institutional authority that can hold P accountable. What if there were another independent punitive authority, such as an independent judiciary, to punish the transgressing P if impeachment fails? In Appendix J, we analyze this setting by introducing another player, the Judiciary (J), who decides whether to rule against a transgressing P with the same punishment, c. We assume that J has her own net incentive to rule against P, which evolves and is subject to conditional revelation, just like C's net incentive to impeach P. We also assume that J is independent, meaning her net incentive to rule against P and C's net incentive to impeach P are mutually independent.

We show that having such J can only partially alleviate the problem presented in Proposition 1. This is because, over time with constant deterrence, the additional threat of punishment from J also becomes as non-credible as a 50/50 coin toss. Thus, in the limit, the overall probability of punishment increases to 0.75, which is higher than 0.5 but still less than one. As a result, although the required threshold for the gain from transgression increases from $a > c$ to $a > 3c$, the challenge in Proposition 1 still exists.

4.4 Many Overlapping Punitive Authorities

As having an additional independent punitive authority can partially address the challenge in Proposition 1, what if there were many such authorities, so that a transgressing P would be punished if at least one of them decides to act against him? This scenario is exemplified by federalism, where the federal government and each state government are considered separate sovereigns. Under the "separate sovereigns" doctrine, even if the transgression is committed only once, it can be addressed multiple times, each time by the judiciary of one of these separate sovereigns.

In Appendix K, we analyze this setting, where a transgressing P is punished if C impeaches him or at least one of many separate judiciaries rules against him. We assume that the net incentive for each judiciary to rule against P evolves and is subject to conditional revelation, just like C's net incentive to impeach P.

We show that, under certain conditions, the challenge in Proposition 1 can be fully addressed in this setting. If all judiciaries and C are independent, meaning that their net incentives to act against P are mutually independent, then, in the limit over time with constant deterrence, the probability of eventual punishment is $1 - 0.5^{N+1}$, where N is the total number of these separate judiciaries. As this number grows, this probability approaches certainty, ensuring that P will comply regardless of the relative size of the punishment and the gain from transgression. In this way, the challenge presented in Proposition 1 can be

overcome with many overlapping punitive authorities.

That said, this solution depends critically on the mutual independence of all overlapping punitive authorities, which is difficult to hold in a polarized political environment. For example, if all these punitive authorities can be categorized into only three independent groups, the probability of punishing a transgressing P in the limit over time with constant deterrence would be $1 - 0.5^3 = 7/8$. Thus, if the gain from transgression is sufficiently large, i.e., for any $a > 7c$, the challenge in Proposition 1 would still exist.

Even if mutual independence were achievable, another concern can arise from having many overlapping punitive authorities. As seen in federalism, the lack of uniformity in state procedures and decisions may lead to political chaos at the federal level, as noted by the Supreme Court of the United States in *Trump v. Anderson* (2024). These limitations highlight the challenges of implementing this potential solution in practice.

4.5 Cultivating Political Culture

In light of these findings, a cultural approach is worth considering, and the literature of constitutional law has been advocating for persistent efforts promoting a political culture against the transgression in question (e.g., Huq, 2018). In our framework, we can model the effect of such efforts as an additional known, positive drift every period, $\mu > 0$, in how the state of the world evolves:

$$w_t = \mu + \rho w_{t-1} + \epsilon_t, \text{ where } \mu > 0,\ 0 < \rho \leq 1, \text{ and } \epsilon_t \sim \mathcal{N}(0, \sigma^2), \text{ i.i.d., with } \sigma > 0. \quad (10)$$

In Appendix L, we show that the cumulative effect of such efforts in the long run would still be too limited to fully address the challenge in Proposition 1, as long as the state of the world is subject to a mean-reverting force ($0 < \rho < 1$), such as a natural decay, however slightly, over time. This is because, as the effect of the effort in each period decays over time, in the limit over time with constant deterrence, the cumulative effect of such efforts would shift the conditional distribution of w_t only finitely, i.e., by $\mu/(1-\rho)$. Thus, the threat of punishment would still resemble a coin toss with a fixed probability, and the challenge in Proposition 1 would still persist.

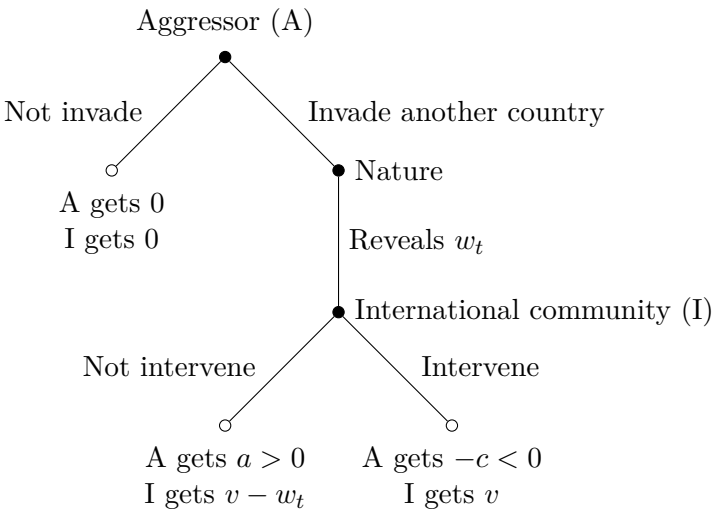
5 Concluding Remarks: Other Applications

In this paper, we model the political nature of punishing transgressions at the apex of government through the conditional revelation of the circumstances that determine whether a transgression will be punished. Our analysis highlights a fundamental challenge in deterring

such transgressions: the very success of deterrence can lead to a lack of new empirical knowledge about the underlying circumstances, gradually eroding the credibility of punishment over time. We demonstrate that several institutional or cultural solutions either fail to fully address the challenge or face inherent limitations.

Beyond transgressions by the head of government, our model also applies to deterring other undesirable behaviors when the exact conditions determining whether such behaviors will get punished only become clear after such behaviors have been committed:

Deterring international aggression. For instance, Figure 4 applies our baseline model to examine whether potential international intervention can deter an aggressor from invading another country. In this context, the conditions determining whether the international community will intervene (w_t) only become clear after the invasion occurs. If these conditions evolve path-dependently yet stochastically, Proposition 1 and our simulation in Figure 3 remain applicable: after a prolonged period of successful deterrence, if the prize of invasion is sufficiently big, the aggressor will eventually invade as the threat of international intervention gradually loses credibility, and consecutive aggressions may follow if the international community does not intervene resolutely ($w_s \leq 0$ or $0 < w_s < \bar{w}$).

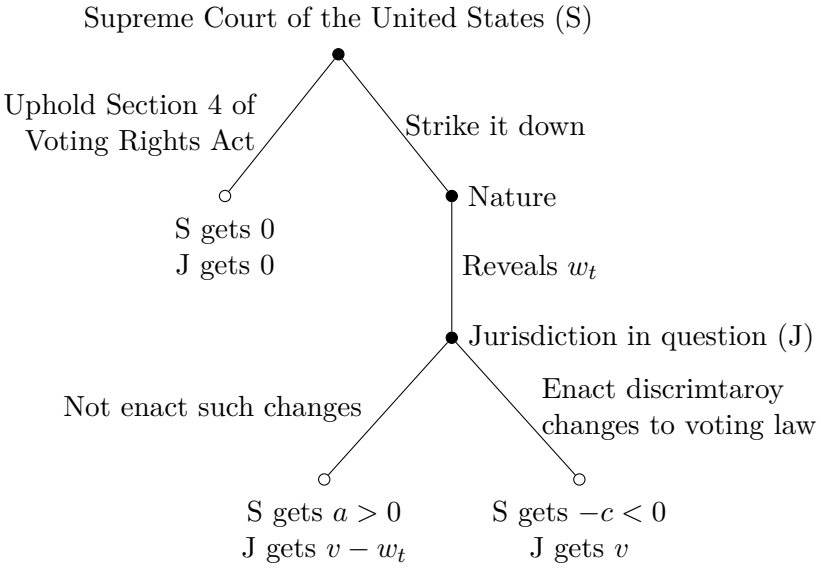


Players know time, $s < t$, and state of the world, w_s, at last invasion.

Figure 4: Deterring international aggression, period t

Beyond deterring undesirable behaviors, the insight from our analysis can shed light on other contexts where the success of a preventive measure may obscure empirical evidence of its necessity, jeopardizing its continuation. We conclude by discussing two examples:

Whether to uphold the Voting Rights Act. First, consider Section 4 of the United States Voting Rights Act of 1965, which provided a formula to identify jurisdictions requiring federal preclearance for changes to voting laws – a measure designed to combat racial discrimination in voting (e.g., Cascio and Washington, 2014). Figure 5 applies our baseline model to this context, where the latent tendency of jurisdictions to enact discriminatory voting laws (w_t) would only become observable if preclearance were removed or rendered ineffective. If this tendency evolves path-dependently yet stochastically, Proposition 1 suggests that the Supreme Court of the United States might eventually perceive preclearance as unnecessary and move to strike it down or significantly weaken it.

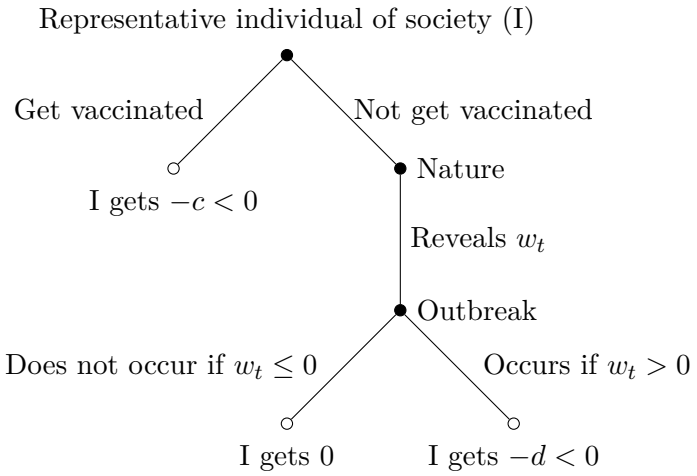


Players know time, $s < t$, and state of the world, w_s, at last time when preclearance provision in Voting Rights Act was ineffective or not in place.

Figure 5: Whether to uphold the Voting Rights Act of 1965, period t

Indeed, in *Shelby County v. Holder* (2013), the Court's majority held that Section 4 of the Voting Rights Act is unconstitutional, arguing that "the conditions that originally justified these measures [of preclearance] no longer characterize voting in the covered jurisdictions." In her dissent, Justice Ginsburg countered that voting equality improvements in these jurisdictions were precisely because "the preclearance remedy [has been] in place in the covered jurisdictions [and] has worked and is continuing to work to stop discriminatory changes." Supporting this view, recent studies have documented negative relative impacts of this ruling on voter registration, turnout, and city-council representation among racial or ethnic minorities (e.g., De Rienzo, 2022; Ricca and Trebbi, 2022; Billings et al., 2024).

Vaccine hesitancy. As another example, the public health literature has long observed that vaccine hesitancy – "delay in acceptance or refusal of vaccination" – can result from vaccination complacency, which occurs when "perceived risks of vaccine-preventable diseases are low and vaccination is not deemed a necessary preventive action" (MacDonald et al., 2015, p. 4161–4162). In particular, "[i]mmunization programme success may, paradoxically, result in complacency and ultimately, hesitancy, as individuals [perceive] the disease the vaccine prevents [as] no longer common" (MacDonald et al., 2015, p. 4163).



I knows time, $s < t$, and state of the world, w_s, at last outbreak in under-vaccinated population.

Figure 6: Whether to get vaccinated, period t

We can formalize this dynamic in Figure 6, adapting our baseline model into a decision problem of a representative individual of society. Here, $c > 0$ represents the vaccination cost, $d > 0$ the potential damage from an outbreak, and w_t the conditions determining whether an outbreak would occur in an under-vaccinated population, which only become observable if the representative individual, and thus the population, choose not to get vaccinated.

When these conditions evolve path-dependently but remain uncertain, Proposition 1's logic applies: prolonged success of vaccination may eventually lead individuals to forgo vaccination as its perceived necessity diminishes. This dynamic, potentially exacerbated by misinformation and conspiracy theories (e.g., Jolley and Douglas, 2014; Martinez-Bravo and Stegmann, 2022; Vergano, 2025), can have severe public health and economic consequences. For instance, Lo and Hotez (2017) warn that "[a] 5% decline in MMR [Measles, Mumps, and Rubella] vaccine coverage in the United States would result in an estimated 3-fold increase in measles cases for children aged 2 to 11 years nationally every year, with an additional $2.1 million in public sector costs."

References

Acemoğlu, Kamer Daron, Nicolás Ajzenman, Cevat Giray Aksoy, Martin Fiszbein, and Carlos Molina. 2025. (Successful) democracies breed their own support. *Review of Economic Studies* **92**: 621–655.

Acemoğlu, Kamer Daron, Georgy Egorov, and Konstantin Sonin. 2013. A political theory of populism. *Quarterly Journal of Economics* **128**: 771–805.

Acemoğlu, Kamer Daron, and James A. Robinson. 2022. Non-modernization: Power–culture trajectories and the dynamics of political institutions. *Annual Review of Political Science* **25**: 323–339.

Agamben, Giorgio. 2005. *State of Exception*. Chicago: University of Chicago Press.

Aristotle. 1998. *Politics*. Indianapolis: Hackett Publishing Company.

Baumgartner, Jody C., and Naoko Kada (eds.). 2003. *Checking Executive Power: Presidential Impeachment in Comparative Perspective*. Westport: Praeger Publishers.

Beccaria, Cesare. 2008. *On Crimes and Punishments and Other Writings*. Toronto: University of Toronto Press.

Becker, Gary S. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* **76**: 169–217.

Besley, Timothy, and Torsten Persson. 2019. Democratic values and institutions. *American Economic Review: Insights* **1**: 59–76.

Bidner, Chris, and Patrick Francois. 2013. The emergence of political accountability. *Quarterly Journal of Economics* **128**: 1397–1448.

Billings, Stephen B., Noah Braun, Daniel B. Jones, and Ying Shi. 2024. Disparate racial impacts of *Shelby County v. Holder* on voter turnout. *Journal of Public Economics* **230**: 105047.

Black, Charles L., Jr., and Philip Bobbitt. 2018. *Impeachment: A Handbook, New Edition*. New Haven: Yale University Press.

Bolton, Patrick, and Howard Rosenthal. 2002. Political intervention in debt contracts. *Journal of Political Economy* **110**: 1103–1134.

Bowman, Frank O., III. 2024. *High Crimes and Misdemeanors: A History of Impeachment for the Age of Trump*. Cambridge: Cambridge University Press, 2nd edition.

Bury, John Bagnell. 1889. *A History of the Later Roman Empire: From Arcadius to Irene (395 A.D. to 800 A.D.), Vol. I*. London: Macmillan.

Cameron, Maxwell A. 1998. Self-coups: Peru, Guatemala, and Russia. *Journal of Democracy*

9: 125–139.

Cascio, Elizabeth U., and Ebonya Washington. 2014. Valuing the vote: The redistribution of voting rights and state funds following the Voting Rights Act of 1965. *Quarterly Journal of Economics* **129**: 379–433.

Chalfin, Aaron, and Justin McCrary. 2017. Criminal deterrence: A review of the literature. *Journal of Economic Literature* **55**: 5–48.

Chiopris, Caterina, Monika Nalepa, and Georg Vanberg. 2021. A wolf in sheep's clothing: Citizen uncertainty and democratic backsliding. Working paper, Harvard University.

Condorcet, Marquis de. 1785. *Essai sur l'Application de l'Analyse à la Probabilité des Déisions Rendus à la Pluralité des Voix*. Paris: Imprimerie Royale.

Congressional Research Service. 2023. *The Constitution of the United States of America: Analysis and Interpretation: Analysis of Cases Decided by the Supreme Court of the United States to June 30, 2022*. Washington: United States Government Publishing Office.

Conte, Marc N., and David L. Kelly. 2021. Understanding the improbable: A survey of fat tails in environmental economics. *Annual Review of Resource Economics* **13**: 289–310.

Crespo, Andrew Manuel. 2019. Impeachment as punishment. *Harvard Law and Policy Review* **13**: 579–592.

De Rienzo, Salvatore M., Jr. 2022. *Shelby County v. Holder* and changes in voting behavior. *American Economist* **67**: 195–210.

Durkheim, Émile. 1973. Two laws of penal evolution. *Economy and Society* **2**: 285–308.

Fearon, James D. 2011. Self-enforcing democracy. *Quarterly Journal of Economics* **126**: 1661–1708.

Ginsburg, Tom, Aziz Z. Huq, and David Landau. 2021. The comparative constitutional law of presidential impeachment. *University of Chicago Law Review* **88**: 81–164.

Gitmez, A. Arda, James A. Robinson, and Mehdi Shadmehr. 2023. Missing discussions: Institutional constraints in the Islamic political tradition. National Bureau of Economic Research Working Paper 30916.

Gratton, Gabriele, and Barton E. Lee. 2024. Liberty, security, and accountability: The rise and fall of illiberal democracies. *Review of Economic Studies* **91**: 340–371.

Grillo, Edoardo, Zhaotian Luo, Monika Nalepa, and Carlo Prato. 2024. Theories of democratic backsliding. *Annual Review of Political Science* **27**: 381–400.

Grillo, Edoardo, and Carlo Prato. 2023. Reference points and democratic backsliding. *American Journal of Political Science* **67**: 71–88.

Grofman, Bernard, Guillermo Owen, and Scott L. Feld. 1983. Thirteen theorems in search

of the truth. *Theory and Decision* **15**: 261–278.

Howell, William G., and Stephane Wolton. 2018. The politician's province. *Quarterly Journal of Political Science* **13**: 119–146.

Huang, Zhangkai, Lixing Li, Guangrong Ma, and Lixin Colin Xu. 2017. Hayek, local information, and commanding heights: Decentralizing state-owned enterprises in China. *American Economic Review* **107**: 2455–2478.

Huq, Aziz Z. 2018. Legal or political checks on apex criminality: An essay on constitutional design. *UCLA Law Review* **65**: 1506–1530.

Jolley, Daniel, and Karen M. Douglas. 2014. The effects of anti-vaccine conspiracy theories on vaccination intentions. *PLoS ONE* **9**: e89177.

Li, Weijia, Gérard Roland, and Yang Xie. 2022. Erosion of state power, corruption control and fiscal capacity. *Economic Journal* **132**: 1542–1565.

Lo, Nathan C., and Peter J. Hotez. 2017. Public health and economic consequences of vaccine hesitancy for measles in the United States. *JAMA Pediatrics* **171**: 887–892.

Luo, Zhaotian, and Adam Przeworski. 2023. Democracy and its vulnerabilities: Dynamics of democratic backsliding. *Quarterly Journal of Political Science* **18**: 105–130.

MacDonald, Noni E., Juhani Eskola, Xiaofeng Liang, Mohuya Chaudhuri, Eve Dubé, Bruce Gellin, Susan Goldstein, Heidi Larson, Mahamane Laouali Manzo, Arthur Reingold, Kinzang Tshering, Yuqing Zhou, Robb Butler, Philippe Duclos, Sherine Guirguis, Ben Hickler, and Melanie Schuster. 2015. Vaccine hesitancy: Definition, scope and determinants. *Vaccine* **33**: 4161–4164.

Madison, James, Jr. 1845. *Debates on the Adoption of the Federal Constitution, with a Diary of the Debates of the Congress of the Confederation, Volume V*. Washington: United States Congress. Revised and arranged by Jonathan Elliot.

Martinez-Bravo, Monica, and Andreas Stegmann. 2022. In vaccines we trust? The effects of the CIA's vaccine ruse on immunization in Pakistan. *Journal of the European Economic Association* **20**: 150–186.

Matějka, Filip, and Guido Tabellini. 2021. Electoral competition with rationally inattentive voters. *Journal of the European Economic Association* **19**: 1899–1935.

Miller, Michael K. 2021. A republic, if you can keep it: Breakdown and erosion in modern democracies. *Journal of Politics* **83**: 198–213.

Monaghan, Chris, Matthew Flinders, and Aziz Z. Huq (eds.). 2024. *Impeachment in a Global Context: Law, Politics, and Comparative Practice*. New York: Routledge.

Myerson, Roger B. 2008. The autocrat's credibility problem and foundations of the consti-

tutional state. *American Political Science Review* **102**: 125–139.

Myerson, Roger B. 2019. The dilemma of presidential impeachment. Perspectives on Economics and Civilization, December 15, 2019.

Nagin, Daniel S. 2013. Deterrence in the twenty-first century. *Crime and Justice* **42**: 199–263.

Ohnesorge, John. 2020. Comparing impeachment regimes. *Duke Journal of Comparative and International Law* **31**: 259–299.

Przeworski, Adam. 2006. Self-enforcing democracy. In Wittman, Donald A., and Barry R. Weingast (eds.) *The Oxford Handbook of Political Economy*. New York: Oxford University Press, 312–328.

Ricca, Federico, and Francesco Trebbi. 2022. Minority underrepresentation in U.S. cities. National Bureau of Economic Research Working Paper 29738.

Rivera, Eduardo, Enrique Seira, and Saumitra Jha. 2025. Apex corruption erodes democratic values. Working paper, Stanford University.

Salzman, Michele Renee. 2021. *The Falls of Rome: Crises, Resilience, and Resurgence in Late Antiquity*. Cambridge: Cambridge University Press.

Schmitt, Carl. 1985. *Political Theology: Four Chapters on the Concept of Sovereignty*. Cambridge: MIT Press.

Stern, Nicholas. 2013. The structure of economic modeling of the potential impacts of climate change: Grafting gross underestimation of risk onto already narrow science models. *Journal of Economic Literature* **51**: 838–859.

Story, Joseph. 1833. *Commentaries on the Constitution of the United States*. Cambridge: Hilliard, Gray, and Company.

Supreme Court of the United States. 2013. Shelby County v. Holder. In *United States Reports*, volume 570. Washington: Supreme Court of the United States, 529–594.

Supreme Court of the United States. 2024. Trump v. Anderson. In *United States Reports*, volume 601. Washington: Supreme Court of the United States.

Svolik, Milan W. 2012. *The Politics of Authoritarian Rule*. Cambridge: Cambridge University Press.

Svolik, Milan W. 2020. When polarization trumps civic virtue: Partisan conflict and the subversion of democracy by incumbents. *Quarterly Journal of Political Science* **15**: 3–31.

Tribe, Laurence H. 1998. Defining "high Crimes and Misdemeanors": Basic principles. *George Washington Law Review* **67**: 712–734.

Tribe, Laurence H., and Joshua Matz. 2018. *To End a Presidency: The Power of Impeach-*

ment. New York: Basic Books.

Vergano, Dan. 2025. The brainwashing campaign that is measles misinformation. Scientific American, April 30, 2025.

von Hayek, Friedrich August. 1945. The use of knowledge in society. *American Economic Review* **35**: 519–530.

Weber, Maximilian K. E. 2004. *The Vocation Lectures*. Indianapolis: Hackett Publishing Company.

Weingast, Barry R. 1997. The political foundations of democracy and the rule of the law. *American Political Science Review* **91**: 245–263.

Weitzman, Martin L. 1974. Prices vs. quantities. *Review of Economic Studies* **41**: 477–491.

Weitzman, Martin L. 2009. On modeling and interpreting the economics of catastrophic climate change. *Review of Economics and Statistics* **91**: 1–19.

Weitzman, Martin L. 2011. Fat-tailed uncertainty in the economics of catastrophic climate change. *Review of Environmental Economics and Policy* **5**: 275–292.

Appendix

A Proof of Proposition 1

Proof. To prove Claim 1, suppose $w_s \leq 0$. C will thus impeach P at period s. By $w_s \leq 0$, $a > c$, and Lemmas 1 and 2, we have

$$\mathbf{P}[w_{s+1} > 0 \mid w_s, s] \leq 0.5 < a/(a+c). \tag{A.1}$$

P will thus transgress at period $s+1$. Claim 1 is proved.

To prove Claim 2, suppose $w_s > 0$. First, C will thus impeach P at period s. Second, at period $s+1$, by the evolution of w_t in Equation (1), P will transgress at period if and only if

$$a/(a+c) > \mathbf{P}[w_{s+1} > 0 \mid w_s, s] = \Phi\left(\rho w_s/\sigma\right), \tag{A.2}$$

or equivalently, if

$$w_s < (\sigma/\rho) \cdot \Phi^{-1}\left(a/\left(a+c\right)\right) \equiv \bar{w}. \tag{A.3}$$

Note $\bar{w} > 0$ because of $a > c$. Claim 2 is thus proved.

To prove Claim 3, suppose $w_s \geq \bar{w}$. By $\bar{w} > 0$, we have $w_s > 0$, and C will thus impeach P at period s. By Claim 2, which is proved above, P will not transgress at period $s+1$. Now consider the periods onward. By $w_s > 0$ and Lemma 1, as time passes with constant deterrence, the conditional probability of punishment decreases and converges to 0.5; by Lemma 2, $0.5 < a/(a+c)$. Therefore, there must exist a period T, when $\mathbf{P}[w_t > 0 \mid w_s, s]$ is so close to 0.5 for the first time that $\mathbf{P}[w_t > 0 \mid w_s, s] < a/(a+c)$ will come to hold. Indeed, in the case of $\rho = 1$, we have

$$T = s + \left\lfloor w_s^2 \Big/ \left(\sigma \cdot \Phi^{-1}\left(a/\left(a+c\right)\right)\right)^2 \right\rfloor + 1, \tag{A.4}$$

where $\lfloor \cdot \rfloor$ is the floor function, and, by $w_s \geq \bar{w}$, we have $T \geq s+2$; in the case of $\rho \in (0,1)$, we have

$$T = s + \left\lfloor \ln\left(\frac{\left(\sigma \cdot \Phi^{-1}\left(a/\left(a+c\right)\right)\right)^2}{w_s^2(1-\rho^2) + \left(\sigma \cdot \Phi^{-1}\left(a/\left(a+c\right)\right)\right)^2} \right) \Big/ 2\ln\rho \right\rfloor + 1, \tag{A.5}$$

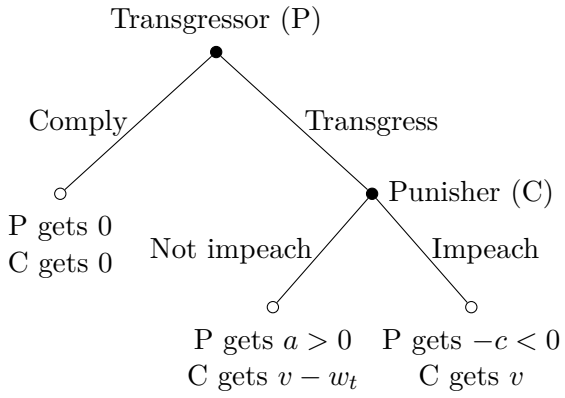
and, by $w_s \geq \bar{w}$, we have $T \geq s+2$, too. P will thus not transgress until period $T-1$ but will transgress at period T. Claim 3 is thus proved, and so is the proposition. □

The proof also implies simple comparative statics of the length of constant deterrence:

Corollary A.1. *T in Proposition 1 increases with w_s and decreases with σ and a/c.*

B Not at the Apex of Government

Figure B.1 shows the structure of the game if the transgression in question were not at the apex of government, but a misconduct or crime in an ordinary setting. Compared with Figure 1, the only difference is that here, without the conditional revelation, the players already know w_t at the beginning of period t. By backward induction, the punisher will punish a transgression if and only if $w_t > 0$; knowing that, the potential transgressor will transgress if and only if $w_t > 0$, independent of the history of transgression and punishment. Therefore, deterrence would not be self-undermining in this setting.



Players know current state of the world, w_t.

Figure B.1: Not at the apex of government, period t

C An Additional Known Component

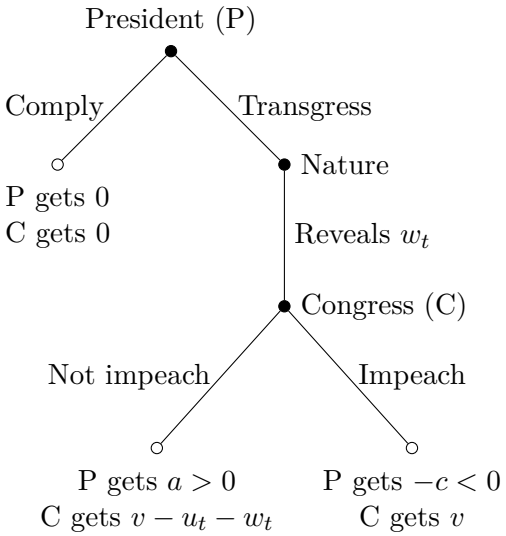
Now consider an alternative setup as in Figure C.1, where the net incentive for C to impeach a transgressing P in period t contains an additional component, u_t, known to the players at the beginning of period t. Since this component captures the "known," i.e., not too wild, side of the state of the world, we further assume that this component is bounded from above, i.e., $u_t \leq \bar{u} < \infty$. Since w_t is still the unknown component, we assume that w_t still follows

$$w_t = \rho w_{t-1} + \epsilon_t, \text{ where } 0 < \rho \leq 1, \text{ and } \epsilon_t \sim \mathcal{N}(0, \sigma^2), \text{ i.i.d., with } \sigma > 0, \tag{C.1}$$

and is independent from u_t.

Given this setup, by backward induction, C will impeach a transgressing P if and only if $w_t > -u_t$. Given this strategy of C, P will transgress if and only if

$$-c \cdot \mathbf{P}[w_t > -u_t | w_s, s, u_t] + a \cdot (1 - \mathbf{P}[w_t > -u_t | w_s, s, u_t]) > 0, \tag{C.2}$$



Players know time, $s < t$, and state of the world, w_s, at last transgression, and u_t.

Figure C.1: An additional known component, period t

or equivalently, if

$$\mathbf{P}[w_t > -u_t | w_s, s, u_t] < a/(a+c). \tag{C.3}$$

Given the evolution of w_t and its independence from u_t, we have

$$(w_t \mid w_s, s, u_t) \sim \begin{cases} \mathcal{N}\left(\rho^{t-s} w_s, \left(1-\rho^{2(t-s)}\right) \cdot \sigma^2 / \left(1-\rho^2\right)\right) & \text{if } \rho \in (0,1), \\ \mathcal{N}\left(w_s, (t-s)\cdot\sigma^2\right) & \text{if } \rho = 1 \end{cases} \tag{C.4}$$

Therefore, we have

$$\mathbf{P}[w_t > -u_t | w_s, s, u_t] = \begin{cases} \Phi\left(\left(\rho^{t-s} w_s + u_t\right) \Big/ \left(\sigma \cdot \sqrt{\frac{1-\rho^{2(t-s)}}{1-\rho^2}}\right)\right) & \text{if } \rho \in (0,1), \\ \Phi\left(\left(w_s + u_t\right) / \left(\sigma \cdot \sqrt{t-s}\right)\right) & \text{if } \rho = 1. \end{cases} \tag{C.5}$$

By $u_t \leq \bar{u} < \infty$, we have, as $t - s \to \infty$,

$$\mathbf{P}[w_t > -u_t | w_s, s, u_t] \leq \begin{cases} \Phi\left(\frac{\rho^{t-s} w_s + \bar{u}}{\sigma \cdot \sqrt{\frac{1-\rho^{2(t-s)}}{1-\rho^2}}}\right) \to \Phi\left(\frac{\bar{u}\cdot\sqrt{1-\rho^2}}{\sigma}\right) < 1 & \text{if } \rho \in (0,1), \\ \Phi\left(\left(w_s + \bar{u}\right) / \left(\sigma \cdot \sqrt{t-s}\right)\right) \to 0.5 & \text{if } \rho = 1. \end{cases} \tag{C.6}$$

Therefore, if $\rho = 1$, whenever $a > c$, perpetual deterrence will be impossible; if $\rho \in (0,1)$, whenever

$$a > c \cdot \Phi\left(\bar{u} \cdot \sqrt{1-\rho^2}/\sigma\right) \Big/ \left(1 - \Phi\left(\bar{u} \cdot \sqrt{1-\rho^2}/\sigma\right)\right) \tag{C.7}$$

perpetual deterrence will be impossible, still. Therefore, the challenge in Proposition 1 will still emerge even if we take into consideration an additional known component of the net incentive for C to impeach a transgressing P.

D Additional Signals from Test Transgressions

Now consider an extension of the baseline model, where we denote the severity of the "real" transgression in question as $\bar{x} \in R$ and the state of the world as $w_t(\bar{x})$. With this notation, the evolution of the state of the world becomes

$$w_t(\bar{x}) = \rho w_{t-1}(\bar{x}) + \epsilon_t, \text{ where } 0 < \rho \leq 1, \text{ and } \epsilon_t \sim \mathcal{N}(0, \sigma^2), \text{ i.i.d., with } \sigma > 0. \tag{D.1}$$

We assume that, at the start of each period t and before the game is played following the baseline setup, P first commits a minor, "test" transgression of severity $\hat{x} < \bar{x}$. This test transgression is costless and generates a public signal of the state of the world,

$$w_t(\hat{x}) = w_t(\bar{x}) + e_t, \text{ where } e_t \sim \mathcal{N}\left(0, (\bar{x} - \hat{x})\varsigma^2\right), \text{ i.i.d., with } \varsigma > 0, \tag{D.2}$$

and all ϵ_ts and e_ts are mutually independent. This signal-generating process implies that the more different the test transgression and the real transgression in question are, the more coarse this signal would be. At one extreme, when $\bar{x} - \hat{x} \to 0$, the real transgression in question is as minor as the test transgression, and the signal from the test transgression would precisely reveal the state of the world; at the other extreme, when $\bar{x} - \hat{x} \to \infty$, the real transgression is extremely different from the test transgression in their nature, and the signal would not reveal much of the state of the world.

We assume that all players will process the signals using Bayesian updating, and we want to understand how these signals may change the dynamics of the game. First, consider the end of period s, when the last real transgression happened with the state of the world revealed as $w_s(\bar{x})$. P's posterior about $w_s(\bar{x})$ is

$$(w_s(\bar{x}) \mid w_s(\bar{x}), s) \sim \mathcal{N}\left(\mu_s^{\text{post}}, 1/p_s^{\text{post}}\right), \tag{D.3}$$

where the mean and precision are, respectively,

$$\mu_s^{\text{post}} = w_s(\bar{x}), \quad p_s^{\text{post}} = \infty. \tag{D.4}$$

Second, at the start of period $s+1$, before receiving the signal, $w_{s+1}(\hat{x})$, while knowing

the evolution of the state of the world, P's prior about $w_{s+1}(\bar{x})$ is

$$(w_{s+1}(\bar{x}) \mid w_s(\bar{x}), s) \sim \mathcal{N}\left(\mu_{s+1}^{\text{prior}}, 1/p_{s+1}^{\text{prior}}\right), \tag{D.5}$$

where the mean and precision are, respectively,

$$\mu_{s+1}^{\text{prior}} = \rho\mu_s^{\text{post}} = \rho w_s(\bar{x}), \quad p_{s+1}^{\text{prior}} = 1/\left(\rho^2/p_s^{\text{post}} + \sigma^2\right) = 1/\sigma^2. \tag{D.6}$$

Since signal follows

$$(w_{s+1}(\hat{x}) \mid w_{s+1}(\bar{x})) \sim \mathcal{N}\left(w_{s+1}(\bar{x}), (\bar{x} - \hat{x})\varsigma^2\right), \tag{D.7}$$

by Bayesian updating, P's posterior about $w_{s+1}(\bar{x})$ is

$$(w_{s+1}(\bar{x}) \mid w_s(\bar{x}), s, w_{s+1}(\hat{x})) \sim \mathcal{N}\left(\mu_{s+1}^{\text{post}}, 1/p_{s+1}^{\text{post}}\right), \tag{D.8}$$

where the mean and precision are, respectively,

$$\mu_{s+1}^{\text{post}} = \frac{p_{s+1}^{\text{prior}} \cdot \mu_{s+1}^{\text{prior}} + w_{s+1}(\hat{x})/(\bar{x} - \hat{x})\varsigma^2}{p_{s+1}^{\text{prior}} + 1/(\bar{x} - \hat{x})\varsigma^2}, \quad p_{s+1}^{\text{post}} = p_{s+1}^{\text{prior}} + 1/(\bar{x} - \hat{x})\varsigma^2. \tag{D.9}$$

We can do this up to the start of period t, assuming that no real transgression has happened from period $s+1$ to $t-1$. At the start of period t, P's prior about $w_t(\bar{x})$ is

$$(w_t(\bar{x}) \mid w_s(\bar{x}), s, w_{s+1}(\hat{x}), \ldots, w_{t-1}(\hat{x})) \sim \mathcal{N}\left(\mu_t^{\text{prior}}, 1/p_t^{\text{prior}}\right), \tag{D.10}$$

where the mean and precision are, respectively,

$$\mu_t^{\text{prior}} = \rho\mu_{t-1}^{\text{post}}, \quad p_t^{\text{prior}} = 1/\left(\rho^2/p_{t-1}^{\text{post}} + \sigma^2\right). \tag{D.11}$$

Since the signal follows

$$(w_t(\hat{x}) \mid w_t(\bar{x})) \sim \mathcal{N}\left(w_t(\bar{x}), (\bar{x} - \hat{x})\varsigma^2\right), \tag{D.12}$$

by Bayesian updating, P's posterior about $w_t(\bar{x})$ is

$$(w_{s+1}(\bar{x}) \mid w_s(\bar{x}), s, w_{s+1}(\hat{x}), \ldots, w_t(\hat{x})) \sim \mathcal{N}\left(\mu_t^{\text{post}}, 1/p_t^{\text{post}}\right), \tag{D.13}$$

where the mean and precision are, respectively,

$$\mu_t^{\text{post}} = \frac{p_t^{\text{prior}} \cdot \mu_t^{\text{prior}} + w_t\left(\hat{x}\right) / \left(\bar{x} - \hat{x}\right) \varsigma^2}{p_t^{\text{prior}} + 1 / \left(\bar{x} - \hat{x}\right) \varsigma^2}, \quad p_t^{\text{post}} = p_t^{\text{prior}} + 1 / \left(\bar{x} - \hat{x}\right) \varsigma^2. \tag{D.14}$$

It is not straightforward to write out the general formulas for μ_t^{post} and p_t^{post}, but we can analyze two extreme cases. If the real transgression in question is as minor as the test transgression, i.e., if $\bar{x} - \hat{x} \to 0$, then we have

$$\mu_t^{\text{prior}} \to \rho w_{t-1}\left(\hat{x}\right), \quad p_t^{\text{prior}} \to 1/\sigma^2, \quad \mu_t^{\text{post}} \to w_t\left(\hat{x}\right), \quad p_t^{\text{post}} \to \infty. \tag{D.15}$$

The game thus converges to the model in Appendix B, which is for misconduct and crimes not at the apex of government, where the current state of the world is known at the start of each period, and the challenge in Proposition 1 would not exist. If the real transgression in question is extremely different from the test transgression, instead, i.e., if $\bar{x} - \hat{x} \to \infty$, then we have

$$\mu_t^{\text{post}} \to \mu_t^{\text{prior}} \to \rho^{t-s} w_s\left(\bar{x}\right), \quad p_t^{\text{post}} \to p_t^{\text{prior}} \to 1 \Big/ \left(\sum_{\tau=0}^{t-s-1} \rho^{2\tau} \right) \sigma^2. \tag{D.16}$$

The game thus converges to the baseline model, and the challenge in Proposition 1 would remain. This demonstrates that the challenge in Proposition 1 arises distinctively when the real transgression in question is not any minor ones, so that little can be learned from these minor ones about the state of the world relevant to the real transgression in question.

E Alternative Ways of Evolution of the World

Deterministic world. If $\sigma = 0$, the current state of the world follows

$$w_t = \rho w_{t-1} = \rho^{t-s} w_s, \text{ where } 0 < \rho \leq 1. \tag{E.1}$$

This implies that the revelation of the state of the world at the last transgression can perfectly indicate the current state, i.e., $w_t > 0$ if and only if $w_s > 0$. Therefore, once a transgression is punished, at each future period, P will understand that any transgression will be punished, so will not transgress. Thus, the challenge in Proposition 1 would not emerge.

Path-independent world. If $\rho = 0$, the current state of the world at each period is independent of past states of the world, i.e.,

$$w_t = \epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right), \text{ i.i.d., with } \sigma > 0. \tag{E.2}$$

The probability of punishment is thus independent of the state of the world at the last transgression, i.e.,

$$\mathbf{P}[w_t > 0 \mid w_s, s] = \mathbf{P}[w_t > 0] = 0.5, \tag{E.3}$$

which is a constant. P will thus always transgress, or always not transgress, in each period, depending on whether or not $a > c$, while independent of the last revelation and how distant it has been in history. Thus, deterrence would not be self-undermining over time.

Oscillatory world. If $\rho < 0$, first observe that

$$(w_{s+1} \mid w_s, s) \sim \mathcal{N}\left(\rho w_s, \sigma^2\right), \text{ where } \sigma > 0 \text{ and } \rho < 0; \tag{E.4}$$

$$(w_{s+2} \mid w_s, s) \sim \mathcal{N}\left(\rho^2 w_s, (1+\rho^2)\sigma^2\right), \text{ where } \sigma > 0 \text{ and } \rho < 0. \tag{E.5}$$

Now we consider two cases. First, for any $w_s > 0$, i.e., right after any impeachment, it must follow that $\rho w_s < 0$ and, thus, $\mathbf{P}[w_{s+1} > 0 \mid w_s, s] < 0.5$. By Lemma 2, this implies that, for any transgression gain that is greater than the punishment, i.e., $a > c$, deterrence will fail for sure at $s+1$.

Second, for any $w_s \leq 0$, instead, i.e., right after any unpunished transgression, it must follow that $\rho^2 w_s \leq 0$ and, thus, $\mathbf{P}[w_{s+2} > 0 \mid w_s, s] \leq 0.5$. By Lemma 2, this implies that, for any transgression gain that is greater than the punishment, i.e., $a > c$, deterrence will fail for sure at $s+2$, if it succeeded at $s+1$. Thus, in an oscillatory world, deterrence would never last, and the challenge in Proposition 1 would have no opportunity to emerge.

Explosive world. If $\rho > 1$, first observe that the conditional distribution is

$$(w_t \mid w_s, s) \sim \mathcal{N}\left(\rho^{t-s} w_s, \left(1-\rho^{2(t-s)}\right) \cdot \sigma^2 / \left(1-\rho^2\right)\right), \text{ where } \sigma > 0 \text{ but } \rho > 1. \tag{E.6}$$

The conditional probability of punishment is thus

$$\mathbf{P}[w_t > 0 \mid w_s, s] = \Phi\left(\rho^{t-s} w_s \Big/ \left(\sigma \cdot \sqrt{\left(1-\rho^{2(t-s)}\right) / \left(1-\rho^2\right)}\right)\right). \tag{E.7}$$

Observe that, for any $t - s \geq 1$ and $w_s > 0$, $\mathbf{P}[w_t > 0 \mid w_s, s]$ decreases with $t - s$ and

$$\mathbf{P}[w_t > 0 \mid w_s, s] \to \Phi\left(\sqrt{\rho^2 - 1} \cdot w_s / \sigma\right) \text{ as } t - s \to \infty. \tag{E.8}$$

Therefore, for any $a > 0$ and $c > 0$, there exists

$$\hat{w} \equiv \left(\sigma/\sqrt{\rho^2 - 1}\right) \cdot \Phi^{-1}\left(a/\left(a + c\right)\right) \tag{E.9}$$

such that, for any $w_s \geq \hat{w}$,

$$\mathbf{P}[w_t > 0 \mid w_s, s] \geq a/(a + c) \tag{E.10}$$

will always hold for any $t - s \geq 1$, i.e., transgression is forever deterred. The challenge in Proposition 1 would thus not exist.

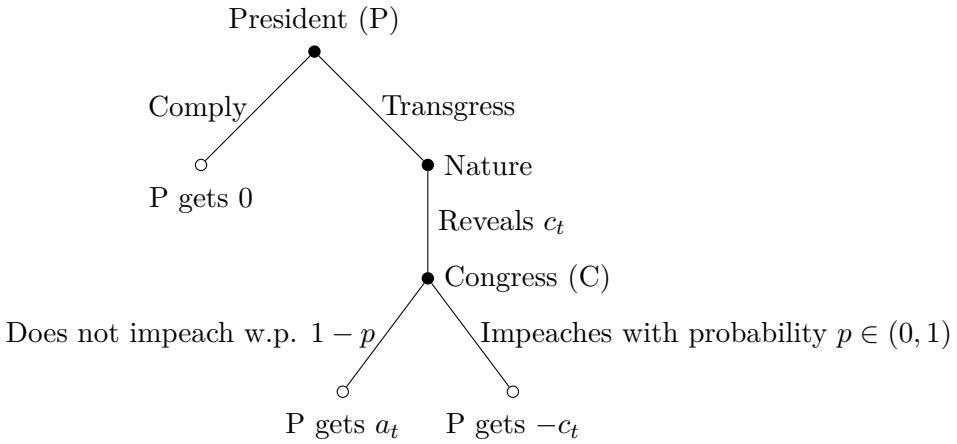
F Payoff Uncertainty

Now consider an alternative setup as in Figure F.1, where strategic uncertainty, i.e., whether C will impeach a transgressing P, is replaced by an exogenous probability of punishment, $p \in (0, 1)$, and payoff uncertainty about the gain and punishment, a_t and c_t, with conditional revelation. About the evolution of (a_t, c_t), we assume that

$$a_t = \rho a_{t-1} + \epsilon_t^a, \quad c_t = \rho c_{t-1} + \epsilon_t^c, \quad \text{where } \rho > 0; \tag{F.1}$$

$$\epsilon_t^a \sim \mathcal{N}(0, \sigma_a^2), \text{ i.i.d., with } \sigma_a \geq 0; \quad \epsilon_t^c \sim \mathcal{N}(0, \sigma_c^2), \text{ i.i.d., with } \sigma_c \geq 0, \tag{F.2}$$

and all ϵ_t^as and ϵ_t^cs are mutually independent. The time of the last transgression is still denoted as $s < t$. We still assume a myopic P.



P knows time, $s < t$, and state of the world, (a_s, c_s), at last transgression.

Figure F.1: Payoff uncertainty, period t

Given this setup, at the beginning of each period t, P will transgress if and only if

$$\mathbf{E}\left[(1-p)a_t - pc_t \mid a_s, c_s, s\right] > 0, \quad \text{i.e.,} \quad (1-p)\cdot\mathbf{E}\left[a_t \mid a_s, c_s, s\right] > p\cdot\mathbf{E}\left[c_t \mid a_s, c_s, s\right]. \tag{F.3}$$

Given the evolution of (a_t, c_t), this condition is equivalent to

$$(1-p)\cdot\rho^{t-s}a_s > p\cdot\rho^{t-s}c_s, \quad \text{i.e.,} \quad (1-p)\cdot a_s > p\cdot c_s \tag{F.4}$$

which is independent of σ_a and σ_c, i.e., the payoff uncertainty, and $t-s$, i.e., how long transgression has been deterred. Therefore, once transgression is deterred, it is forever deterred. We see that the challenge in Proposition 1 is about strategic uncertainty, not payoff uncertainty.

G Forward-looking Players

In the main text, we have assumed that the players are myopic. In this section, we show that the challenge presented in Proposition 1 is robust with respect to forward-looking players. In particular, we assume that, once a P is impeached, a new P will replace him immediately, and the exiting P will suffer a permanent punishment of $c > 0$ every period afterward; all the players are forward-looking with an infinite horizon, and the discount factor is $\beta \in (0,1)$.

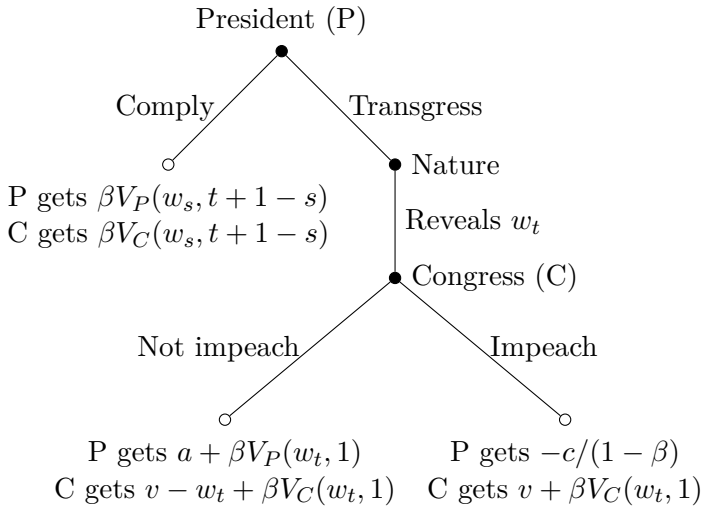
We consider symmetric, pure-strategy Markov-perfect equilibria (SMPE), where "symmetric" means that all Ps are restricted to adopting the same strategy. The payoff-relevant variables, i.e., state variables, are thus only the last revelation of the state of the world, i.e., at the last regression, w_s, and how long time it has been since the last transgression, $t-s$. We thus write the net present values that the players enjoy in equilibrium at the beginning of period t as $V_i(w_s, t-s)$, $i = C, P$.

Figure G.1 then illustrates the structure of the subgame starting from t, taking the continuation value in equilibrium for the future period as given. At the beginning of period t, the payoff-relevant variables, w_s and $t-s$, are known to all the players. As in Section 2, for simplicity, we still assume that C will not impeach P if indifferent between impeaching and not impeaching, and P will not transgress if indifferent between transgressing and complying.

We first establish the existence and uniqueness of the equilibrium. By Figure G.1, we can pin down C's equilibrium strategy, since she cannot affect the future state of the world:

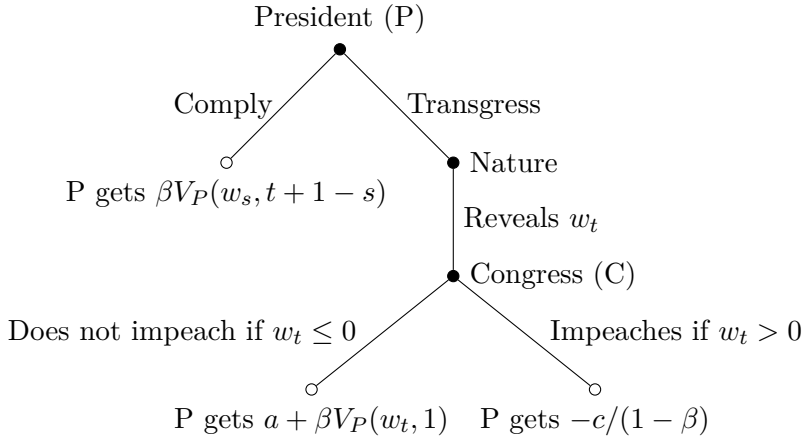
Lemma G.1. *In any SMPE, C's strategy is to impeach any transgressing P at period t if and only if $w_t > 0$.*

Lemma G.1 helps us reduce the subgame at period t, as in Figure G.1, into a decision



Players know time, $s < t$, and state of the world, w_s, at last transgression.

Figure G.1: Forward-looking players and permanent punishment, subgame at period t



Players know time, $s < t$, and state of the world, w_s, at last transgression.

Figure G.2: Forward-looking players and permanent punishment, subgame at t reduced

problem for P, as in Figure G.2. P's strategy in any SMPE is thus to transgress at period t if and only if

$$\begin{aligned}&\mathbf{E}\left[a + \beta V_P(w_t, 1) \mid w_t \leq 0, w_s, s\right] \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s] - (c/(1-\beta)) \cdot \mathbf{P}[w_t > 0 \mid w_s, s] \\ &> \beta \cdot V_P(w_s, t+1-s), \end{aligned} \tag{G.1}$$

or equivalently, if

$$\begin{aligned}&a \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s] - (c/(1-\beta)) \cdot \mathbf{P}[w_t > 0 \mid w_s, s] \\ &+ \beta \cdot \mathbf{E}\left[V_P(w_t, 1) \mid w_t \leq 0, w_s, s\right] \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s] > \beta \cdot V_P(w_s, t+1-s), \end{aligned} \tag{G.2}$$

The Bellman equation for P at the beginning of period t is thus

$$\begin{aligned} V_P(w_s, t-s) = \max \big\{ & a \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s] - (c/(1-\beta)) \cdot \mathbf{P}[w_t > 0 \mid w_s, s] \\ & + \beta \cdot \mathbf{E}\left[V_P(w_t, 1) \mid w_t \leq 0, w_s, s\right] \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s], \\ & \beta \cdot V_P(w_s, t+1-s) \big\}. \end{aligned} \tag{G.3}$$

We can then establish the existence and uniqueness of the equilibrium:

Proposition G.1. *There exists a unique SMPE, in which C will impeach any transgressing P at period t if and only if $w_t > 0$, and P's strategy solves the Bellman equation,*

$$\begin{aligned} V_P(w_s, t-s) = \max \big\{ & a \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s] - (c/(1-\beta)) \cdot \mathbf{P}[w_t > 0 \mid w_s, s] \\ & + \beta \cdot \mathbf{E}\left[V_P(w_t, 1) \mid w_t \leq 0, w_s, s\right] \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s], \\ & \beta \cdot V_P(w_s, t+1-s) \big\}. \end{aligned} \tag{G.4}$$

Proof. First, by Lemma G.1, we have C's strategy in any SMPE, if exists. Second, observe that any $V_P(\cdot, \cdot)$ in equilibrium, if exists, must be bounded by $[0, a/(1-\beta)]$. This is because, an always-complying strategy will generate a zero net present value for P, while the best possible scenario for P would not be better than an always-transgressing strategy being never punished, which would generate a net present value of $a/(1-\beta)$ for P.

Now denote the set of functions from $\mathbb{R} \times \mathbb{N}^+$ to $\mathbb{R}$ that are bounded by $[0, a/(1-\beta)]$ as

$$S \equiv \left\{ \nu : \mathbb{R} \times \mathbb{N}^+ \to \mathbb{R}, \nu(\cdot, \cdot) \in [0, a/\left(1-\beta\right)] \right\}, \tag{G.5}$$

and endow it with the sup-norm metrics,

$$\rho(\nu, \mu) = \sup_{(x,y) \in \mathbb{R} \times \mathbb{N}^+} |\nu(x, y) - \mu(x, y)|. \tag{G.6}$$

Also denote the operator from S to S that is based on the Bellman equation as

$$\begin{aligned} G(\nu)(w_s, t-s) \equiv \max \big\{ & a \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s] - (c/(1-\beta)) \cdot \mathbf{P}[w_t > 0 \mid w_s, s] \\ & + \beta \cdot \mathbf{E}\left[\nu(w_t, 1) \mid w_t \leq 0, w_s, s\right] \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s], \\ & \beta \cdot \nu(w_s, t+1-s) \big\}. \end{aligned} \tag{G.7}$$

It is easy to show that G satisfies Blackwell's sufficient conditions for a contraction of modules $\beta \in (0, 1)$. By Banach's contraction mapping theorem, G thus admits a unique fixed point, which is the value function $V_P(\cdot, \cdot)$ in the unique SMPE. The proposition is thus proved. □

We can then show a result parallel to Proposition 1:

Proposition G.2. *Given any $a > c/(1-\beta)$, there does not exist a finite level of the state of the world at a transgression at period s, $w_s < \infty$, such that, in the unique SMPE, P will always comply from period $s+1$ onward.*

Proof. Suppose that there exist such $w_s < \infty$ that, in the unique SMPE, P will comply in any period $t \geq s+1$. Since P will comply in any period $t \geq s+1$, then we have, for any $t \geq s+1$,

$$\begin{aligned} &a \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s] - (c/(1-\beta)) \cdot \mathbf{P}[w_t > 0 \mid w_s, s] \\ &+ \beta \cdot \mathbf{E}\left[V_P(w_t, 1) \mid w_t \leq 0, w_s, s\right] \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s] \leq \beta \cdot V_P(w_s, t+1-s), \end{aligned} \tag{G.8}$$

where

$$V_P(w_s, t+1-s) = 0. \tag{G.9}$$

This condition thus becomes

$$\begin{aligned} &a \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s] - (c/(1-\beta)) \cdot \mathbf{P}[w_t > 0 \mid w_s, s] \\ &+ \beta \cdot \mathbf{E}\left[V_P(w_t, 1) \mid w_t \leq 0, w_s, s\right] \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s] \leq 0, \end{aligned} \tag{G.10}$$

or equivalently,

$$\begin{aligned} &a - (a + (c/(1-\beta))) \cdot \mathbf{P}[w_t > 0 \mid w_s, s] \\ &+ \beta \cdot \mathbf{E}\left[V_P(w_t, 1) \mid w_t \leq 0, w_s, s\right] \cdot \mathbf{P}[w_t \leq 0 \mid w_s, s] \leq 0. \end{aligned} \tag{G.11}$$

Note that, by Lemma 1, $\mathbf{P}[w_t > 0 \mid w_s, s] \to 0.5$ as $t - s \to \infty$; we also know $V_P(\cdot, \cdot) \in [0, a/(1-\beta)]$. Therefore, by $a > c/(1-\beta)$, as $t - s \to \infty$, the left-hand side of Inequation (G.11) is strictly positive. This inequation thus cannot hold as $t - s \to \infty$, contradicting to what we have supposed. The proposition is thus proved by contradiction. □

H Severe Punishment

Now we turn to the case where the punishment for transgression is sufficiently severe, i.e., $c \geq a$. Lemmas 1 and 2 imply that perpetual deterrence will become possible, in that once transgression is deterred, it will be forever deterred:

Proposition H.1. *For any $a \leq c$, we have $\bar{w} \equiv (\sigma/\rho) \cdot \Phi^{-1}(a/(a+c)) \leq 0$, and*

1. *if $w_s < \bar{w}$, then C will not impeach P at period s, and P will transgress at $s+1$;*
2. *if $\bar{w} \leq w_s \leq 0$, then C will not impeach P at s, but P will never transgress from $s+1$ onward;*

3. if $w_s > 0$, then C will impeach P at s, and P will never transgress from $s+1$ onward.

Proof. About Claim 3, since $w_s > 0$, C will impeach P at period s. By Lemma 1, the conditional probability of punishment at period $s+1$ is greater than 0.5; by $a \leq c$ and Lemma 2, even a 50/50 coin toss is able to deter transgression. Therefore, P will not transgress at period $s+1$. Since no transgression has happened since period $s+1$, period s was still the last time when the state of the world was revealed. The same argument applied to period $s+1$ thus applies from period $s+2$ onward, and P will thus not transgress from period $s+2$ onward. Claim 3 is thus proved.

About Claim 2, by $w_s \leq 0$, C will not impeach P at period s. About period $s+1$, as in the proof of Proposition 1, P will transgress if and only if

$$w_s < (\sigma/\rho) \cdot \Phi^{-1}\left(a/(a+c)\right) \equiv \bar{w}, \tag{H.1}$$

though, by $a \leq c$, $\bar{w} \leq 0$. Therefore, if $\bar{w} \leq w_s \leq 0$, then the threat to punish is sufficiently credible to deter P from transgressing. About period $s+2$, by the fact that P will not transgress at period $s+1$, period s is still the last time when the state of the world was revealed. By $w_s \leq 0$ and Lemma 1, that implies that the threat of punishment at period $s+2$ is even more credible than that at period $s+1$, so it is sufficiently credible to deter P from transgressing at period $s+2$. The same argument applies from period $s+2$ onward. Claim 2 is thus proved.

About Claim 1, since $w_s < \bar{w}$, P will transgress at period $s+1$; since $\bar{w} \leq 0$, $w_s < 0$ and C will thus not impeach P at period s. Claim 1 and the proposition are thus proved. □

Proposition H.1 suggests that the self-undermining nature of deterrence of transgressions will not be a problem for perpetual deterrence, if the punishment is sufficiently severe. The relative size of P's gain and punishment for transgression is thus a key parameter about the institution surrounding the players. A question follows: if someone cannot observe the institution directly, can she find it out from the dynamics of transgression and impeachment?

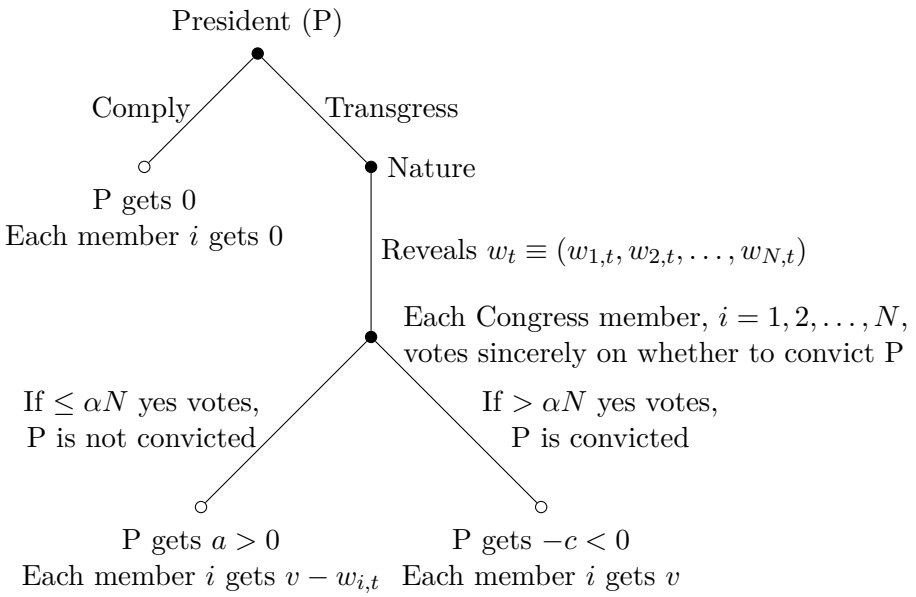
Corollary H.1. *Only observing a transgressing P being impeached at period $s < t$ and constant deterrence from then to the current period t, one cannot identify whether or not the institution will make the constant deterrence forever, i.e., whether $a \leq c$ or $a > c$; she can tell $a > c$ once another transgression happens.*

The intuition behind Corollary H.1 is simple: an ongoing period of constant deterrence can happen in both the cases of $a > c$ and $a \leq c$. In this sense, only observing an ongoing period of constant deterrence will not tell us much about the underlying institution. Therefore, not only does constant deterrence generate a lack of new empirical knowledge

about the state of the world, as contained in the setup of our model, but it cannot tell much about the institution by itself; in order to learn about the institution, one has to wait for the institution to be under stress, i.e., when a transgression happens.

I Congress of Many Non-partisan Members

Now consider an extension as in Figure I.1. The main difference from Section 2 is that now Congress consists of $N \geq 3$ members, each denoted by $i = 1, 2, \ldots, N$. With this Congress of multiple members, after P transgresses and Nature reveals the state of the world, each Congress member will vote sincerely, i.e., considering oneself to be pivotal when voting, on whether to convict P. P will then be convicted if and only if there are greater than αN yes votes, where $\alpha \in [0, 1)$ is exogenous and represents the voting rule for conviction.



Players know time, $s < t$, and state of the world, w_s, at last transgression.

Figure I.1: Congress of $N \geq 3$ members, with voting rule $\alpha \in [0, 1)$, period t

All the other features of the extension follow the model in Section 2, with accommodations necessary for a Congress of multiple members. About the state of the world, we assume that each Congress member faces her own net incentive to convict a transgressing P, $w_{i,t}$. The state of the world is thus a vector now, $w_t \equiv (w_{1,t}, w_{2,t}, \ldots, w_{N,t})$. We assume that each component of the state of the world evolves as modeled in Section 2, i.e., for any i, $w_{i,t} = \rho_i w_{i,t-1} + \epsilon_{i,t}$, where $0 < \rho_i \leq 1$ and $\epsilon_{i,t} \sim \mathcal{N}(0, \sigma_i^2)$, independent and identically distributed, with $\sigma_i > 0$. We further assume that all $w_{i,t}$s are mutually independent across all i, so that all the Congress members are non-partisan.

Given this setting, at each period t, by backward induction, each Congress member i will vote to convict a transgressing P if and only if $w_{i,t} > 0$, while P will transgress if and only if the threat of conviction is not credible enough, i.e.,

$$\mathbf{P}[W_t > \alpha N \mid w_s, s] < a/(a+c), \tag{I.1}$$

where W_t is the number of convicting votes against a transgressing P.

Limit over time with constant deterrence. About this threat of conviction, since all $w_{i,t}$s are mutually independent across all i, Lemma 1 still applies. In particular, as time passes with constant deterrence, the threat from each Congress member i to vote to convict P, conditional on the time and state of the world at the last transgression, s and w_s, will converge to a 50/50 coin toss, i.e.,

$$\mathbf{P}[w_{i,t} > 0 \mid w_s, s] \to 0.5 \text{ as } t - s \to \infty. \tag{I.2}$$

Therefore, in this limit, the conditional distribution of the number of votes to convict a transgressing P will converge to a Binomial distribution, with the number of Bernoulli trials being N and the success probability being 0.5, i.e.,

$$(W_t \mid w_s, s) \xrightarrow{d} W \sim \mathcal{B}(N, 0.5) \text{ as } t - s \to \infty. \tag{I.3}$$

By the Weak Law of Large Numbers, this implies that, in this limit over time with constant deterrence, as the size of Congress, N, increases, it will be almost certain that the proportion of the Congress members voting to convict a transgressing P will be close to one half, i.e.,

$$W/N \xrightarrow{p} 0.5 \text{ as } N \to \infty. \tag{I.4}$$

Applying the voting rule to this limit, we have the following proposition:

Proposition I.1. *For any transgression gain and punishment for P, given any super-majority rule for conviction in impeachment, in the limit over time with constant deterrence, as the size of Congress increases, it will be almost certain that Congress will not convict any transgressing P and, therefore, P will transgress; given any minority rule, in the limit over time with constant deterrence, as the size of Congress increases, it will be almost certain that Congress will convict any transgressing P and, therefore, P will comply. Mathematically, for*

any $a > 0$ *and* $c > 0$,

$$\begin{aligned} &\textit{if } 0.5 < \alpha < 1, \textit{ then } \mathbf{P}\left[W > \alpha N\right] \to 0 < a/(a+c) \textit{ as } N \to \infty; \\ &\textit{if } 0 \leq \alpha < 0.5, \textit{ then } \mathbf{P}\left[W > \alpha N\right] \to 1 \geq a/(a+c) \textit{ as } N \to \infty. \end{aligned} \tag{I.5}$$

Proposition I.1 suggests that admitting many non-partisan members to Congress, given any super-majority rule for conviction, will make the problem presented in Proposition 1 even worse: it is almost certain that transgression cannot be forever deterred, even if the transgression gain is smaller than the punishment, i.e., $a \leq c$.

J Another Call by an Independent Judiciary

Now consider another extension as in Figure J.1. The main difference from Section 2 is that now, in case that a transgressing P is not impeached by C, a Judiciary (J) can make another call on the matter. If J rules against P, then P will still be punished; if J rules in favor of P, then P will receive the gain from transgression. The net incentive for J to rule against P is w'_t, where we assume that

$$w'_t = \rho' w'_{t-1} + \epsilon'_t, \text{ where } 0 < \rho' \leq 1, \text{ and } \epsilon'_t \sim \mathcal{N}(0, \sigma'^2), \text{ i.i.d., with } \sigma' > 0, \tag{J.1}$$

and is subject to the same conditional revelation as w_t. We also assume that the judiciary is independent, such that w'_t and w_t are mutually independent.

Given this setting, P will transgress if and only if

$$\mathbf{P}\left[w_t > 0 \text{ or } w'_t > 0 \mid w_s, w'_s, s\right] < a/(a+c). \tag{J.2}$$

Note that, by the mutual independence between w'_t and w_t, we have

$$\mathbf{P}\left[w_t > 0 \text{ or } w'_t > 0 \mid w_s, w'_s, s\right] = 1 - \mathbf{P}\left[w_t \leq 0 \mid w_s, s\right] \cdot \mathbf{P}\left[w'_t \leq 0 \mid w'_s, s\right]. \tag{J.3}$$

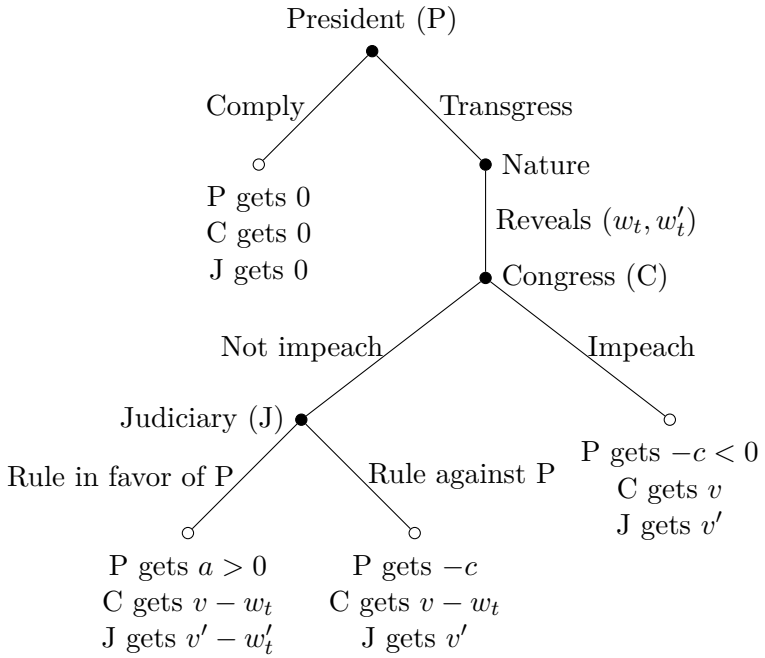
As $t - s \to \infty$, we have

$$\mathbf{P}\left[w_t > 0 \text{ or } w'_t > 0 \mid w_s, w'_s, s\right] \to 1 - 0.5 \times 0.5 = 0.75. \tag{J.4}$$

Therefore, if

$$a/(a+c) > 0.75, \text{ i.e., } a > 3c, \tag{J.5}$$

then constant deterrence cannot last forever. Therefore, the challenge in Proposition 1 will



Players know time, $s < t$, and state of the world, (w_s, w'_s), at last transgression.

Figure J.1: Another call by judiciary, period t

still emerge, even though the required threshold for the transgression gain relative to the punishment is higher, i.e., $a > 3c$, instead of $a > c$.

K Federalism

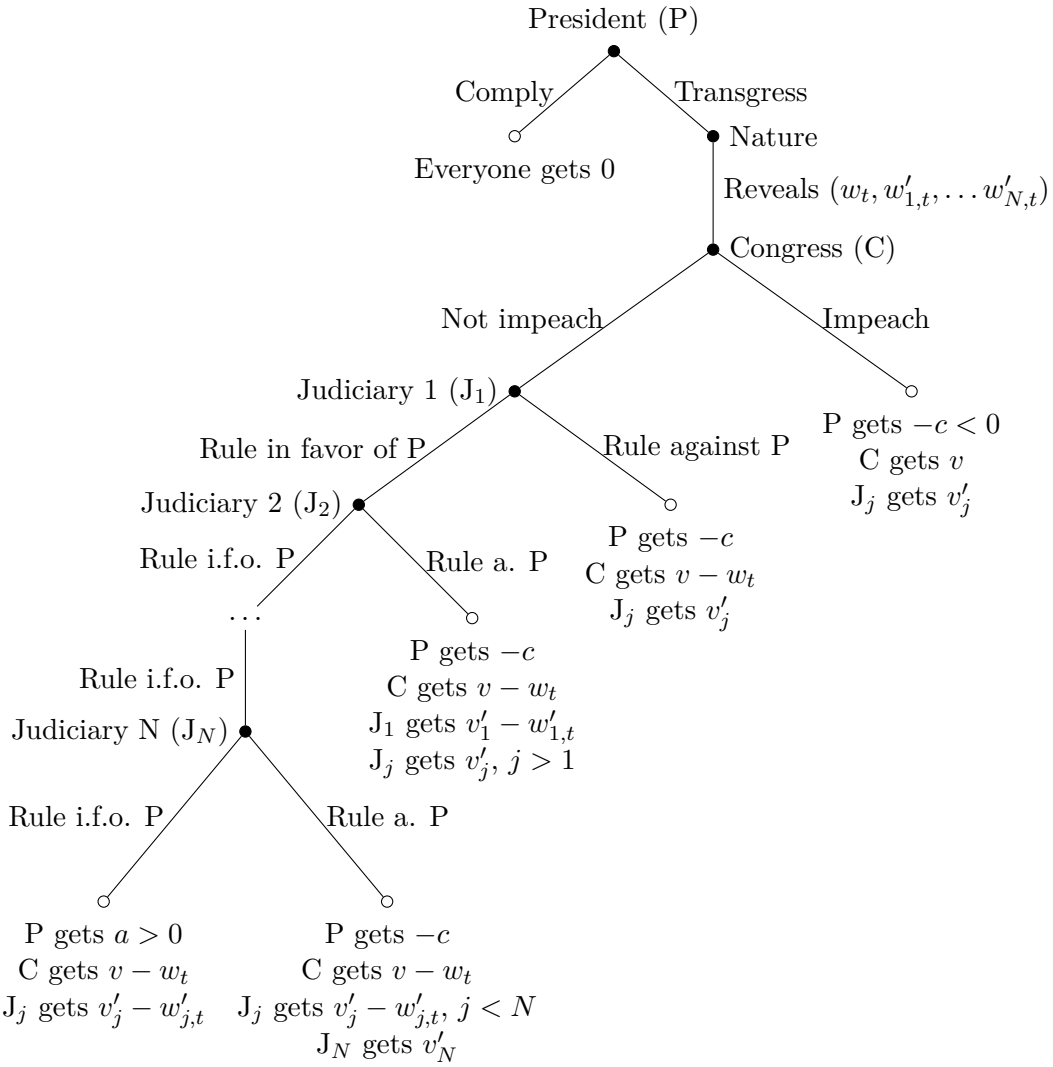
Now consider a further extension as in Figure K.1. The main difference from Section J is now, each of N Judiciaries, J_i, $i = 1, \ldots, N$, in sequence, can make her call if the transgressing P is not punished by C and the judiciaries before her, such that a transgressing P will be punished if C impeaches him, or if at least one J_i rules against him. The net incentive for each J_i to rule against the transgressing P is $w'_{i,t}$ and subject to conditional revelation.

Now suppose that all judiciaries are independent from C and each other, i.e., for each $i = 1, \ldots, N$,

$$w'_{i,t} = \rho'_i w'_{i,t-1} + \epsilon'_{i,t}, \text{ where } 0 < \rho'_i \leq 1, \text{ and } \epsilon'_{i,t} \sim \mathcal{N}(0, \sigma'^2_i), \text{ i.i.d., with } \sigma'_i > 0, \qquad \text{(K.1)}$$

and $w'_{1,t}$, ..., $w'_{N,t}$, and w_t are mutually independent. Thus, P will transgress if and only if

$$\mathbf{P}\left[w_t > 0 \text{ or } w'_{1,t} > 0 \text{ or } \ldots \text{ or } w'_{N,t} > 0 \mid w_s, w'_{1,s}, \ldots, w'_{N,s}, s\right] < a/(a+c). \qquad \text{(K.2)}$$



Players know time, $s < t$, and state of the world, $(w_s, w'_{1,s}, \dots w'_{N,s})$, at last transgression.

Figure K.1: Federalism, period t

Note that, by the mutual independence among $w'_{1,t}$, ..., $w'_{N,t}$, and w_t, we have

$$\begin{aligned}&\mathbf{P}\left[w_t > 0 \text{ or } w'_{1,t} > 0 \text{ or } \dots \text{ or } w'_{N,t} > 0 \mid w_s, w'_{1,s}, \dots, w'_{N,s}, s\right] \\ &= 1 - \mathbf{P}\left[w_t \leq 0 \mid w_s, s\right] \cdot \prod_{i=1}^{N} \mathbf{P}\left[w'_{i,t} \leq 0 \mid w'_{i,s}, s\right]. \end{aligned} \tag{K.3}$$

As $t - s \to \infty$, we have

$$\mathbf{P}\left[w_t > 0 \text{ or } w'_{1,t} > 0 \text{ or } \dots \text{ or } w'_{N,t} > 0 \mid w_s, w'_{1,s}, \dots, w'_{N,s}, s\right] \to 1 - 0.5^{N+1}. \tag{K.4}$$

In this limit, as $N \to \infty$, we have

$$\mathbf{P}\left[w_t > 0 \text{ or } w'_{1,t} > 0 \text{ or } \ldots \text{ or } w'_{N,t} > 0 \mid w_s, w'_{1,s}, \ldots, w'_{N,s}, s\right] \to 1. \tag{K.5}$$

Therefore, for any given $a > 0$ and $c > 0$, in the limit over time with constant deterrence, as the number of independent judiciaries increases, it will be almost certain that a transgressing P will be punished and, therefore, P will comply. In this sense, the challenge in Proposition 1 can be overcome asymptotically.

Note that this result is conditional on the mutual independence of all the judiciaries and C. To see this, suppose instead that all the judiciaries face the same incentive as C's, i.e., for any $i = 1, \ldots, N$, $w'_{i,t} = w_t$. This setting is then reduced to the baseline model, and Proposition 1 holds.

L A Positive Drift

Now suppose that the state of the world evolves with a positive drift every period, i.e.,

$$w_t = \mu + \rho w_{t-1} + \epsilon_t, \text{ where } \mu > 0,\ 0 < \rho \leq 1, \text{ and } \epsilon_t \sim \mathcal{N}(0, \sigma^2), \text{ i.i.d., with } \sigma > 0. \tag{L.1}$$

Its distribution, conditional on its value and the time at the last transgression, is thus

$$(w_t \mid w_s, s) \sim \begin{cases} \mathcal{N}\left(\frac{1-\rho^{t-s}}{1-\rho} \cdot \mu + \rho^{t-s} w_s, \frac{1-\rho^{2(t-s)}}{1-\rho^2} \cdot \sigma^2\right) & \text{if } \rho \in (0,1), \\ \mathcal{N}\left((t-s)\mu + w_s, (t-s) \cdot \sigma^2\right) & \text{if } \rho = 1. \end{cases} \tag{L.2}$$

The conditional probability of punishment is thus

$$\mathbf{P}[w_t > 0 \mid w_s, s] = \begin{cases} \Phi\left(\left(\frac{1-\rho^{t-s}}{1-\rho} \cdot \mu + \rho^{t-s} w_s\right) \Big/ \left(\sigma \cdot \sqrt{\frac{1-\rho^{2(t-s)}}{1-\rho^2}}\right)\right) & \text{if } \rho \in (0,1), \\ \Phi\left((\mu(t-s) + w_s) \,/\, \left(\sigma \cdot \sqrt{t-s}\right)\right) & \text{if } \rho = 1. \end{cases} \tag{L.3}$$

Therefore, as $t - s \to \infty$, we have

$$\mathbf{P}[w_t > 0 \mid w_s, s] \to \begin{cases} \Phi\left((\mu/\sigma) \cdot \sqrt{(1+\rho)/(1-\rho)}\right) \in (0.5, 1) & \text{if } \rho \in (0,1), \\ 1 & \text{if } \rho = 1. \end{cases} \tag{L.4}$$

Therefore, having a constant positive drift can only partially alleviate the problem presented in Proposition 1, if the effect of historical shocks decays over time; it cannot overcome the challenge, unless there is no such decay.