When Consensus Acquits

Capture and the design of the Sanhedrin's capital procedure

Torun Dewan

London School of Economics

Abstract

Courts can be captured: a sovereign, a prosecutor, or a powerful litigant who wants a conviction can sweep or suborn the bench, and a procedure for rendering judgment must reckon with the possibility that its own judges are not free. We model the Sanhedrin's capital procedure as a design for exactly this problem. Its centrepiece is a rule that looks perverse – a court that convicts unanimously acquits the accused (Sanhedrin 17a) – and we show it is the optimal response to capture: among all the rules a court could commit to, it minimises error against a captor who best-responds. A captured bench manufactures consensus, so unanimity is the signature of capture, and a rule that refuses to convict on it deters the manipulation the court cannot observe. The rest of the procedure follows from the same primitive – the pro-acquittal asymmetry, junior-first voting, and benches that grow with the gravity of the charge – and the account sorts anti-capture institutions by the threat they face.

1 Introduction

Courts can be captured. A power that wants a conviction can pack a bench, suborn it, or sweep it away. A captured bench returns the verdict it is given. The instrument of capture is agreement. A compromised court does not deliberate; it concurs. So the verdict capture can force is the unanimous one. A procedure for collective judgment must guard against this, yet cannot see it. It observes the votes, not whether they were freely cast. The benches it most needs to fear are those that agree most readily: an assembly that ratifies by acclamation, a politburo that records no dissent. Each manufactures the consensus that, taken at face value, would most compel us.

The response is to read agreement against itself. If capture manufactures consensus, then consensus is the signature of capture. A court that would deter what it cannot observe must refuse the verdict capture forces. An ancient court built exactly this refusal into its law. The Sanhedrin was the capital court of Talmudic procedure. It tried the gravest charges before benches of twenty-three or seventy-one. It tilted every step toward acquittal. And it laid down a rule that looks perverse: a bench that convicts unanimously thereby acquits the accused (*Sanhedri she-ra'u kullan le-chova – potrin oto*, "a Sanhedrin all of whom saw fit to convict, they acquit him", Sanhedrin 17a).[1] The most agreement of all produces no conviction. We argue that this rule is what a court designs when it may be captured. The unanimous verdict is the one blunt capture forces. So it is the one the court will not honour. Cheap manipulation buys only a verdict the court discards; the routes to any other verdict are dearer, as we show. So capture is never worth attempting.

Consensus is self-discrediting. In any body whose agreement can be manufactured, agreement is evidence against itself. Where a verdict can be produced without deliberation, the free bench and the captured one are observationally close precisely at unanimity, so an optimal rule conditions acquittal on the consensus it can no longer trust, and conviction on the bare dissent it can. The Sanhedrin is one institution that implements this rule; the logic binds wherever judgment is collective and capture is possible.

The formal study of collective verdicts has, by contrast, assumed the difficulty away. Its standard treatment takes the judges' *signals* to be conditionally independent and their votes informative; the count of convicting votes is then a sufficient statistic for guilt, the posterior is increasing in it, and the optimal rule is a threshold – convict when the count is high enough (Condorcet, 1785). Its strategic heirs sharpened this: Austen-Smith and Banks (1996) and Feddersen and Pesendorfer (1998) asked when voting one's signal is itself an equilibrium, and found

[1] The acquittal bias runs through the procedure. The Mishnah needs a bare majority to acquit but a majority of two to convict, opens each trial for the defence, and lets a judge revise his vote toward acquittal but not toward conviction, all of which the same mishna sets opposite monetary procedure, where a bare majority of one decides either way (Sanhedrin 4:1–4:2); the asymmetry is itself read from Scripture – *lo tihyeh aharei rabim le-ra'ot*, "do not follow a multitude to do evil" (Exodus 23:2; Sanhedrin 2a) – so the differential cost of the two errors enters the law as a verse, not a modelling choice.

that unanimity rules can fail it. More agreement is always more reason to convict. Conditional independence of the judges' signals is what delivers this monotonicity – the premise all these treatments share, and the one we drop – and that independence is exactly what capture destroys. Restore the possibility that the bench is not free, in the one direction a captor pushes it, and the threshold inverts at the top: the count stays decisive until the last vote, at which agreement turns from the strongest evidence into the weakest.

The mechanism is this. Suppose that with some probability the bench is compromised – deferring to a dominant member, or directed to its verdict by a power that wants a conviction – and that a compromised bench returns unanimous conviction whatever the truth. The compromise we model is *state-independent*: it loads on consensus regardless of guilt. This is the clean limit of, but not identical to, an informational cascade (Banerjee, 1992; Bikhchandani et al., 1992), whose direction is set by early signals and so remains correlated with the truth: a cascade would leave even the unanimous count informative about guilt, and so would not deliver the clean inversion of Proposition 1. What we require is the sharper, state-independent case, in which the consensus carries no information about the state at all. The designer sees the votes, not the regime. Then unanimity is no longer strong evidence of guilt. It is the outcome a compromised bench manufactures, so observing it shifts weight onto the captured regime, under which the vote said nothing. The count does not become uninformative – but it is discounted, and once capture is likely enough it falls below a lone dissent. A single dissent, by contrast, is almost impossible under capture, so it certifies that the court was deliberating and that all but one of its independent judgments pointed to guilt. The lone dissenter does not weaken the case. He authenticates it. Remove him and one can no longer tell a convinced court from a captured one.

We then let the prospect of capture be a choice rather than an accident. A patron who would convict the innocent can corrupt a bench, but corruption can only manufacture the consensus the rule refuses to honour, so the rule deters the manipulation it cannot observe. Its force is a commitment, which is why the procedure is fixed in advance and not revisable in the case at hand; and because courts are pressed toward conviction, the rule is asymmetric in just the way the sources are, withholding conviction from a unanimous bench but never imposing it.

The same primitive organises the rest of the procedure. We read five of its features – the pro-acquittal asymmetry, the acquittal on unanimity, the order in which judges vote, the graded benches, and the strategic restraint of the judges themselves – as one response to a single problem: how to render judgment when the court may not be free. The procedure manufactures the independence it then audits, sizes itself to price capture out, and commits in advance to a rule it would be tempted to break, so that the manipulation it guards against is never attempted.

Reading a procedure of the Talmud as a designed response to a strategic problem places this paper within a small program of such readings, after Aumann and Maschler (1985), which take a sugya not as a solution to a problem the analyst poses but as an institution – precisely legislated,

defended in argument, and built against a strategic problem of information, commitment, or division. The mechanism we put to work is not itself new: that a state-independent component loading on consensus drives the likelihood ratio at the top toward one, and can thereby invert the posterior in the count, is the sharp limit of the correlated voting that already blunts aggregation in the Condorcet setting (Ladha, 1992). What is new is the object to which it is turned, and three claims about it. The first is a dialogue with the canonical result. Feddersen and Pesendorfer (1998) showed that requiring unanimity to convict is a *bad* rule, one that convicts the innocent. We do not impose unanimity and find it wanting; we observe the unanimous verdict as an event on a majority-conviction bench, and reverse its *meaning*: what a conviction rule would read as the strongest case becomes the weakest, and the optimal procedure acquits on it. The second is that the contamination is *one-sided*: it loads on the convict-consensus alone, so the rule's suspicion is asymmetric and tracks the direction in which courts are captured, and it is this, not the non-monotonicity as such, that parts the argument from the correlated-votes literature. The third is that capture can be deterred by design: a court that commits to withhold conviction from the verdict capture can cheaply force prices out the manipulation it cannot observe, and that commitment is the loss-minimising rule against a captor who best-responds, from which a cluster of the Sanhedrin's features – its pro-acquittal asymmetry, its graded benches, its order of voting – then follow. Because the rule answers one kind of capture and not another, the theory sorts anti-capture institutions by the threat they face: a consensus-discrediting rule where a guaranteed verdict against a marked defendant must be deterred, the Sanhedrin or the procedure one would build against the show trial, and its absence where capture works through a standing majority, as in a packed constitutional court.

2 Capture, aggregation, and commitment

The reading offered here belongs to a program that takes the procedures of the Talmud not as antecedents to be admired but as designed institutions – legislated against strategic problems and defended in argument – after Aumann and Maschler (1985). Read so, a procedure built against the capture of a court speaks to three problems of modern political economy, each with a literature that has met it apart from the others. We set the connections out not to claim the procedure anticipates these literatures but to mark where its single primitive – a doubt about whether a deliberating body is free – touches each.

Consensus as the signature of capture. The rule's governing idea is that agreement which can be manufactured is evidence against itself. A literature on unfree institutions rests on the same observation: there the visible agreement is the artefact, not the fact. Hegemonic parties manufacture majorities far larger than they need (Magaloni, 2006); governments manipulate elections by margins well beyond winning, to transmit an image of invincibility that shapes the conduct of rivals and citizens alike (Simpser, 2013); and a public that falsifies its preferences

sustains a consensus which conceals private dissent and can collapse without warning (Kuran, 1991). The institutions of dictatorship – assemblies, parties, courts – are read in this work as devices that organise the appearance of assent (Gandhi, 2008; Svolik, 2012). The Sanhedrin's rule is the response a designer makes once this is grasped: it declines to read manufactured agreement as agreement, and so refuses to convict on the very consensus an unfree bench exists to supply. The court-specific form of the threat – a bench staffed for loyalty rather than independence – is the capture our model takes as its primitive (Nalepa, 2022; Chiopris et al., 2025).

Aggregation when independence is in doubt. The graded benches and the non-monotone reading of the count are a claim about information. The Condorcet tradition and its strategic heirs aggregate *conditionally independent* signals, under which more agreement is always more reason to convict; our departure is to ask what a body should infer when that independence itself is uncertain, and to find the count then read non-monotonically and discounted at the top. The bridge to that literature – set out where we relate the argument to the theory of juries below – is the dependence capture introduces: a common cause that moves the whole bench at once is the sharp limit of the correlated voting that already blunts aggregation in the Condorcet setting. The graded bench does double work in consequence, enlarged where error is gravest both to price out capture and to aggregate more signal at once.

Commitment and the rigidity of rules. Under sincere voting the deterrent binds only if the court can hold itself to acquitting a unanimous bench it would, in the case at hand, rather convict (under strategic voting the acquittal is ex-post optimal and the question does not arise; we return to this below). The procedure's answer – a verdict fixed in advance as a mechanical function of the count, revisable only by a greater court – is an instance of the founding result in the theory of commitment, that rules attain what discretion cannot (Kydland and Prescott 1977). The same logic carries a political economy of institutions as commitments: constitutions that bind a sovereign by raising the price of reneging (North and Weingast, 1989), and an independent judiciary understood not as a neutral arbiter but as the device through which competing powers enforce mutual restraint across time (Landes and Posner, 1975; Stephenson, 2003). The rigidity of codified procedure, which an administrative eye reads as mere inflexibility, is on this account the source of the deterrent: a rule that cannot be set aside in the individual case is precisely the rule a captor cannot hope to see bent.

3 Model

A defendant is guilty G or innocent I; the prior is $\pi = \Pr(G)$. A bench of n judges returns a profile of votes, and the court observes only the number of convicting votes $k \in \{0, \dots, n\}$.

With probability $1 - \lambda$ the bench is *free* (we reserve "deliberative" for the communication models of the jury literature): its judges deliberate honestly, receiving conditionally indepen-

dent signals and voting informatively, convicting with probability p if G and $1-p$ if I, where $p > \frac{1}{2}$. With probability λ the bench is *compromised* and returns unanimous conviction, $k = n$, regardless of the state. The regime is unobserved and independent of guilt. (The compromised regime stands in for any state-independent failure of independence that concentrates on consensus; unanimous conviction is the sharp case.) The rate λ is a property of the polity in which the court sits – the standing pressure of sovereigns and factions toward a manufactured verdict – and the procedure is built to confront it, not assumed to abolish it; we return to its equilibrium determination in Section 7 and find it stays positive.

Compromise is a property of the bench, not a fact each judge reads off his own seat: a judge knows only that *he himself* is free, and what he infers about his colleagues' votes he derives from the regime probabilities rather than assumes. When the only compromise is full, this inference is trivial: a judge who knows he himself is free has thereby ruled out a fully compromised bench, on which he too would have been instructed, and so need not read the count to know his colleagues' votes informative. It ceases to be trivial once a third, *partially* compromised regime is admitted – a faction holding all but one seat, introduced with the captor's problem in Section 7 – which makes the lone free judge's inference the heart of the partial-capture analysis. The designer sees only the count, not the regime, and so conditions on different information than the judges do.

The costs of error are asymmetric. Convicting the innocent is the graver error, summarised by a threshold τ on the likelihood ratio: the verdict convicts at count k iff $L(k) \equiv \Pr(k \mid G)/\Pr(k \mid I) \geq \tau$, with τ large.

4 Consensus is self-discrediting

For $k < n$ only the free regime can produce the profile, so

$$L(k) = \left(\frac{p}{1-p}\right)^{2k-n}, \qquad \text{strictly increasing in } k. \tag{1}$$

At unanimity both regimes contribute:

$$L(n) = \frac{(1-\lambda)\,p^n + \lambda}{(1-\lambda)(1-p)^n + \lambda}. \tag{2}$$

Proposition 1 (Non-monotonicity)**.** *$L(n)$ is continuous and strictly decreasing in λ, with $L(n) = \left(\frac{p}{1-p}\right)^n$ at $\lambda = 0$ and $L(n) \to 1$ as $\lambda \to 1$. Since $L(n-1) = \left(\frac{p}{1-p}\right)^{n-2}$ does not depend on λ, there is a unique $\lambda^\star \in (0,1)$ with $L(n) = L(n-1)$, and for $\lambda > \lambda^\star$ the posterior probability of guilt is non-monotone in the conviction count: it rises to a peak at $k = n-1$ and falls at $k = n$, so $L(n) < L(n-1)$.*

Proof. For $k < n$ the compromised regime contributes nothing, so $\Pr(k \mid \omega) = (1-\lambda)\binom{n}{k}\rho_\omega^k(1-\rho_\omega)^{n-k}$ with $\rho_G = p$ and $\rho_I = 1-p$; the binomial coefficients cancel, giving $L(k) = (p/(1-p))^{2k-n}$,

strictly increasing because $p > \frac{1}{2}$, and in particular $L(n-1) = (p/(1-p))^{n-2}$. At $k = n$ both regimes contribute: $\Pr(n \mid G) = (1-\lambda)p^n + \lambda$ and $\Pr(n \mid I) = (1-\lambda)(1-p)^n + \lambda$ are affine in λ, so $L(n)$ is a ratio of affine functions and its derivative has the constant sign of $ad - bc$ with $a = 1-p^n$, $c = p^n$, $b = 1-(1-p)^n$, $d = (1-p)^n$; here $ad-bc = (1-p)^n - p^n < 0$, so $L(n)$ is strictly decreasing on $[0,1]$, from $(p/(1-p))^n$ at $\lambda = 0$ to 1 at $\lambda = 1$. As $1 < L(n-1) < (p/(1-p))^n$ for $n \geq 3$ – and a capital bench has $n \geq 23$, so the bound binds with room to spare – the equation $L(n) = L(n-1)$ has a unique root $\lambda^\star \in (0,1)$, with $L(n) < L(n-1)$ for $\lambda > \lambda^\star$. (For $n = 2$ the construction degenerates, since then $L(n-1) = L(1) = 1 = \lim_{\lambda \to 1} L(n)$ and no interior root exists; the capital benches we study have $n \geq 23$, so this is moot.) Since L is increasing on $\{0, \ldots, n-1\}$, the likelihood ratio then peaks at $k = n-1$. □

Proposition 2 (Unanimous conviction acquits)**.** *For $\lambda > \lambda^\star$ and any threshold $\tau \in \big(L(n),\, L(n-1)\big)$, the optimal verdict convicts at $k = n-1$ and acquits at $k = n$. The conviction set is an interval $[\,k_\tau,\, n-1\,]$ that excludes unanimity.*

Proof. With prior π and asymmetric error costs, the Bayes-optimal verdict convicts at count k iff the posterior odds $\frac{\pi}{1-\pi} L(k)$ exceed the cost ratio, that is iff $L(k) \geq \tau$ for the implied threshold τ. On $\{0, \ldots, n-1\}$ the ratio L is strictly increasing, so $\{k \leq n-1 : L(k) \geq \tau\} = [k_\tau, n-1]$ with $k_\tau = \min\{k : L(k) \geq \tau\}$, nonempty since $\tau < L(n-1)$. At $k = n$, $\tau > L(n)$ gives $L(n) < \tau$, so the verdict acquits. The conviction set is thus $[k_\tau, n-1]$, excluding unanimity. □

Remark 1 (Robustness to the reading of the sugya). Propositions 1 and 2 turn on one feature of the count and nothing else: that capture loads a state-independent mass on consensus, $k = n$. They do not ask *when* the count is taken. The sugya admits more than one reading of which consensus the rule audits – the Bavli's *ra'u kullan le-chova*, "all were of the view to convict" (Sanhedrin 17a), reads as the bench's settled position, while Maimonides fixes it at the opening, a bench that *opens* the case already united for conviction (*she-patchu kullam... techilah*, Hilchot Sanhedrin 9:1). Our foundation is neutral between them: whether k records the opening tally or the final one, a captured bench loads the same atom on unanimity and the same non-monotone discount follows. We do not adjudicate the reading, because the result we build on holds under either. The strategic results of the next section are more committed – they read the count as the bench's sequential, junior-first expression of position – and we mark that commitment where it is incurred.

5 Reading

The rule conditions the verdict on the vote *and* on its credibility. A near-unanimous bench with one dissent is the most incriminating outcome the procedure admits, because dissent is the mark of a free court. Unanimity is exactly what a captured court produces, so the law refuses to convict on it. This is the sense the codifiers give the acquittal. Maimonides states it as settled

law: a Sanhedrin all of whom *open* the capital case by convicting – *she-patchu kullam. . . techilah ve-amru kullan chayav* – acquits, for a capital bench must hold among it those who would argue the defendant's merit, and here there were none (*Mishneh Torah*, Hilchot Sanhedrin 9:1). The rule passes from the Bavli (Rav Kahana, Sanhedrin 17a) into the code but travels no further: the *Shulchan Arukh* does not carry capital procedure at all, the courts that administered it having lapsed with ordination, so the forward chain ends at Maimonides rather than at the later codes – the law is preserved as doctrine, not as a live docket. Maimonides locates the trigger at the opening, in the absence of a defender; the model locates it in the consensus count. By the remark above we need not choose between them: under either, a free bench is far likelier to break ranks – to produce the dissent, or in Maimonides' telling the defender, that a captured bench cannot. Free unanimity is not impossible, only rare, arising with probability p^n on a guilty bench; that small positive chance is precisely what the inference at unanimity weighs against capture. Unanimity is otherwise left as the signature of a bench that could not break ranks, and the absence indicts the process, not the man.

The other features of the procedure fall into place around this. Beginning the vote from the junior judges (Sanhedrin 4:2) is the device that protects the independence the rule audits: the least senior speak before they can defer, which keeps the free regime informative and holds λ down. The graded benches and the majority of two to convict set the asymmetric price of error that the threshold τ records. The rule that consensus acquits is the audit; junior-first is the safeguard that makes the audit meaningful.

6 Strategic voting

The baseline takes judges to vote informatively. We now let them vote strategically. Judges share an interest in a correct verdict but weigh the two errors asymmetrically, convicting only when their posterior on guilt clears a high threshold, and each conditions his vote on the event in which it is decisive. We ask whether the rule of Proposition 2 survives strategic voting. It does, and is sharpened.

One reading is now fixed. The foundation left open *when* the count is taken; the strategic argument does not – it reads the count as the bench's sequential, junior-first expression of position (Sanhedrin 4:2, 36a), a tally assembled through an ordered and observed vote. The pivotal logic that follows lives on that process and is claimed under that reading. Its *outcome* – that a free bench bent on conviction does not complete the unanimity – survives the opening reading as well, where a convinced judge opens for the defence rather than complete the lockstep; but the mechanism below is stated for the sequential count, and that is the one interpretive commitment the rest of the paper makes.

The rule introduces a decisive event at the top of the count: a judge whose vote moves the bench between $k = n - 1$ and $k = n$. There, voting to convict makes the bench unanimous and

acquits, while dissenting holds the count at $n-1$ and *convicts*. The pivotal judge's incentive is thus inverted at the top – he dissents to convict, and concurs to acquit.

We solve this sequential game. Judges vote in turn, the most junior first, each seeing the votes already cast (Sanhedrin 4:2); that the order is observed is not an auxiliary assumption but the premise of the law's own reason for it – that juniors voting after the greatest would be swayed by him, "neither shall you answer after the Master" (Sanhedrin 36a). The following lemma is the engine for all that follows.

Lemma 1 (A free bench's verdict is its decisive judge's choice)**.** *Under the rule of Proposition 2 and sequential voting with observed history, take any free-bench history at which a judge moves with the count at $n-1$, so that his vote completes the unanimity that acquits or withholds it and convicts. Because the order is observed he knows his vote is decisive between conviction ($k=n-1$) and acquittal ($k=n$), so sequential rationality has him cast it for the verdict he prefers under his posterior – whatever that posterior, and however he reads the prefix: he dissents if he would convict, concurs if he would acquit. Every path by which a free bench's count reaches n runs through such a decisive vote, so a free bench completes the unanimity only when its decisive judge prefers acquittal, which the rule then grants; a free bench that would convict stops instead at $k=n-1$. An observed $k=n$ therefore never warrants conviction – it is capture's manufacture or a free bench's own route to acquittal – and the argument fixes this without computing the pivot posterior or assuming the prefix informative.*

Proof. At the stated history the judge's two votes lead, by the rule, to distinct verdicts: convicting makes the count n and acquits, dissenting holds it at $n-1$ and convicts. Since the history is observed he knows which vote yields which verdict, so his choice is between securing conviction and securing acquittal, and sequential rationality selects the action whose verdict he weakly prefers given his information. He dissents whenever he would convict on his posterior and concurs whenever he would acquit – a step that uses only the decisiveness of his vote, not any reading of the $n-1$ prior votes. A count reaches n only when some judge moving at $n-1$ votes to convict; by the foregoing that judge preferred acquittal, which the rule grants. Hence a free bench that prefers conviction never completes the unanimity but convicts at $n-1$, and a free bench that reaches $k=n$ is one the rule rightly acquits; a compromised bench returns $k=n$ by construction, irrespective of any judge's information. (The judge reasons on his own posterior given a free bench, not on the designer's λ-mixed $L(n)$ of Proposition 1: the designer sees only the count and mixes over the regime, whereas a free judge knows his bench is free.) □

The lemma reasons at one node; the next proposition shows that node is not a fiction. It is reached by an equilibrium of the sequential game, and the action the lemma reads off it is the one every equilibrium forces there – so the keystone is an equilibrium statement, not an artefact of a convenient selection.

Proposition 3 (The rule's outcome holds in every equilibrium)**.** *Take the junior-first sequential game under the rule of Proposition 2, with $\tau < L(n-1)$ so a count of $n-1$ convicts. In every sequential equilibrium the senior, the last to move, casts the decisive vote of Lemma 1 at the top pivot: reaching a history of $n-1$ prior convictions, he dissents when his posterior favours conviction and concurs when it favours acquittal. Hence in every sequential equilibrium a free bench returns $k = n$ only through a last mover who prefers acquittal, while a free bench whose last mover prefers conviction secures it short of unanimity, at a count in $[k_\tau, n-1]$; the rule turns no intended conviction into the acquittal at unanimity. The vote tracks the senior's posterior, and so his signal, so the equilibrium is responsive – and one exists.*

Proof. A finite game of incomplete information has a sequential equilibrium (Kreps and Wilson, 1982); fix one. Consider any history – on the path or off it – at which the senior moves with $n-1$ convictions already cast. The order is observed, so he knows that convicting completes the unanimity and acquits, while dissenting holds the count at $n-1$, which the rule convicts because $n-1 \in [k_\tau, n-1]$ when $\tau < L(n-1)$. His two actions lead to distinct verdicts, and sequential rationality requires the one his posterior weakly prefers: dissent if he would convict, concur if he would acquit. This is forced at every such information set, being the uniquely sequentially-rational choice at a decisive and observed node, and it invokes neither the value of the pivot posterior nor any reading of the prior votes – exactly the content of Lemma 1, now pinned to an equilibrium rather than asserted of a vote in isolation. A free bench reaches $k = n$ only through this completing vote, cast only when the senior prefers acquittal, so a free bench's unanimity records a last mover who preferred it, which the rule grants; a free bench whose last mover prefers conviction has him dissent and convict at $n-1$. A compromised bench returns $k = n$ by construction and is acquitted. The senior's vote tracks his signal – the equilibrium is *responsive* – exactly when his two signal-contingent posteriors straddle the threshold, $\frac{\pi}{1-\pi} L_{\text{prefix}} \frac{p}{1-p} \geq \tau > \frac{\pi}{1-\pi} L_{\text{prefix}} \frac{1-p}{p}$, with L_{prefix} the likelihood ratio he reads off the $n-1$ history. On a non-degenerate set of priors and thresholds this holds, and there a responsive sequential equilibrium exists. Where it fails the senior is unresponsive at the pivot, but the outcome-level conclusion of Lemma 1 is unaffected: a free convicting bench still stops at $k = n-1$. We claim no more of the interior – not that the prior votes are informative, nor that the equilibrium is unique. □

Proposition 4 (The rule binds only on captured benches)**.** *Under the rule of Proposition 2 and strategic voting, the acquittal at unanimity reverses no free-bench conviction. A free bench that would convict casts its decisive vote for dissent and obtains its conviction at a count in $[k_\tau, n-1]$ (Lemma 1); a unanimous count from a free bench reflects a decisive judge who preferred acquittal, which the rule grants. The only conviction the rule withholds at unanimity is the one a captured bench, forced to $k = n$ irrespective of the judges' information, manufactures.*

Proof. Immediate from Proposition 3, restated in the language of capture. In every sequential

equilibrium a free bench that prefers conviction has its last mover dissent and convict at $n-1$, where the verdict is the increasing threshold of Proposition 2; a free bench that reaches $k = n$ does so only through a last mover who preferred acquittal, so the rule's acquittal there reverses no conviction it sought. A compromised bench returns $k = n$ by construction, independent of any judge's information, and is acquitted. Hence the only verdict the rule overturns against a free bench's intent is the manufactured unanimity of capture. □

Corollary 1 (The top pivot's incentive is reversed)**.** *The two rules turn the pivotal incentive at the top in opposite directions. Under the rule that requires unanimity* to convict *(Feddersen and Pesendorfer, 1998), a juror is decisive only when all others convict; conditioning on that event raises his posterior on guilt, so a juror with an innocent signal may rationally convict. Under the rule that* acquits *on unanimity, the decisive completing vote instead triggers acquittal, so a convinced free judge dissents rather than cast it (Lemma 1). The comparison is exact only at this single top pivot and for a finite bench. We claim nothing mirroring the substance of Feddersen–Pesendorfer, whose result is asymptotic – the probability of convicting the innocent stays bounded away from zero as the jury grows – and rests on a full equilibrium characterisation; here there is one pivot, no large-bench limit, and no welfare comparison. The two rules share the pivotal logic and reverse its direction; that is all the corollary asserts.*

The question of who casts the dissent now settles, though more loosely than a single designated judge. A bench approaching unanimity faces a last possible dissent at the top pivot, and by Lemma 1 whoever stands there secures the conviction by dissenting if he would convict; only a decisive judge who instead prefers acquittal completes the unanimity, and the rule then acquits as he intends. Under junior-first (Sanhedrin 4:2) the most senior judge is the last mover and so the backstop: if a convicting consensus reaches him unbroken and he too would convict, he breaks it. A judge already convinced earlier in the sequence may pre-empt him, dissenting before the bench arrives at $n-1$; the keystone does not turn on which of them does it, only that a free bench bent on conviction does not convict through unanimity (the next remark). Junior-first earns its place here for a separate and surer reason – it keeps the early signals independent by denying deference its opening – and the two rules interlock on that: junior-first manufactures the independence that unanimity-acquits audits.

Remark 2 (The order is not innocuous). For the monotone, symmetric rules Dekel and Piccione (2000) study, the informative equilibria of sequential and simultaneous voting coincide, and the order of voting is irrelevant to what the bench learns. Their irrelevance is established for those rules; the acquit-at-unanimity rule is *non-monotone* – it withholds conviction at the very top of the count – and there is no reason to expect the irrelevance to survive it, so the order becomes payoff-relevant. This is why the procedure troubles to fix an order at all: junior-first both keeps the early signals independent, by denying deference its opening, and places the residual top pivot on the senior. Order does work here that, for the monotone rules of the jury literature, it cannot.

Remark 3 (An outcome-level argument). The lemma asks nothing of the equilibrium beyond sequential rationality at a decisive vote. It does not compute the pivot posterior, does not require the $n-1$ prior votes to be informative, and does not characterise the full sequential game – in which a convinced judge may dissent pre-emptively and interior votes near the top need not be informative. It uses only that the last vote at the top is decisive between conviction and acquittal, a fact of the rule and the observed order; the conclusion therefore holds across responsive equilibria and under either sincere or strategic voting. What it does *not* claim is that a free bench never reaches $k = n$: a decisive judge who prefers acquittal secures it by completing the unanimity, and there the rule acquits as he intends. The content is that the rule never converts a free bench's intended conviction into an acquittal – which is all the deterrence argument requires. Proposition 3 sharpens this from an outcome read off one vote into an equilibrium statement: the decisive node is reached by an equilibrium of the sequential game, the action there is the one *every* equilibrium forces, and a responsive equilibrium exists, so the lemma is no artefact of a convenient selection. What we still decline to claim is what the deterrent does not need and the jury literature does not deliver – that the interior votes are informative, or that the equilibrium is unique.

7 Capture and the asymmetry of the rule

So far the probability λ that the bench is compromised has been exogenous. We now let it be chosen. A *patron* – a sovereign, a prosecutor, a powerful litigant – gains $B > 0$ from *securing* the conviction of this defendant whatever the truth, where B is the incremental gain over the verdict a free bench would return of its own accord, so that abstaining from capture is normalised to a payoff of zero; he may *capture* the bench at cost $\kappa < B$. Capture is *blunt*: it manufactures agreement, returning unanimous conviction $k = n$. To manufacture instead a particular interior count – in particular the lone dissent that would leave $k = n - 1$ – costs a further δ, because a captured bench must orchestrate a holdout it has no reason to seat. This increment is not a bare assumption; its sign and direction are supplied by the economics of vote-buying, as the following remark records. What the deterrence below asks of it is only a magnitude – that δ be large enough that $B < \kappa + \delta$, so that producing consensus is cheap where producing *credible disagreement* is dear. It is the same asymmetry Proposition 3 exploited – a free bench dissents of its own accord, a captured one cannot. The designer commits to a verdict rule before the patron moves.

Remark 4 (The price of a margin short of unanimity, after Groseclose and Snyder). The sign of δ is what the economics of vote-buying predicts. Let the patron face a counter-buyer – a defendant, a faction – who moves last and may bid one vote back, and recall that one vote acquits. A unanimous convicting bench offers no pivot to peel; a count short of it leaves a margin of one to defend against that last move. This is the Groseclose and Snyder (1996) configuration:

a coalition a counter-buyer can attack is held only by overbuying toward a robust supermajority. Under a margin that acquits on one vote the only attack-proof convicting coalition is the whole bench, so the captor is driven toward the very unanimity the rule refuses, and any count short of it carries a supermajority premium $\delta > 0$, larger the narrower the acquittal margin.

Proposition 5 (Capture deterrence)**.** *Fix the standing compromise rate $\lambda > \lambda^\star$ of Section 7 above its exogenous floor. Under any rule that convicts at $k = n$, the consensus a compromised bench manufactures is itself the conviction the patron seeks, worth buying whenever $B > \kappa$, so the standing rate translates one-for-one into* capture-induced wrongful convictions. *Under the rule of Proposition 2, which acquits at $k = n$, that manufactured consensus delivers acquittal, and converting it into a conviction – faking a dissent to reach $k = n - 1$ – costs the further δ, with $\kappa + \delta > B$. The patron cannot profitably turn the standing consensus into a verdict. The rule does not drive λ to zero – it cannot, and Proposition 1 needs $\lambda > \lambda^\star$ for the rule to bind at all – but it drives the* capture-induced wrongful-conviction rate *to zero: the consensus a captor can force is acquitted, and the verdict he would need is priced out. This is deterrence, not manipulation-proofness: it holds under the cost wedge $B < \kappa + \delta$, where manufacturing consensus is cheap and manufacturing a credible dissent is dear – the asymmetry the procedure builds, and which Proposition 2 turns against the captor.*

Proof. If the rule convicts at $k = n$, the consensus a compromised bench manufactures is a conviction worth B at cost κ, profitable whenever $B > \kappa$, and the standing rate λ then passes straight into wrongful convictions. Under the rule of Proposition 2 the patron's routes are blunt capture, which yields $k = n$ and so acquittal at net $-\kappa$, and capture with a fabricated dissent, which yields $k = n - 1$ and conviction at net $B - \kappa - \delta$. With δ large, $B < \kappa + \delta$, both are loss-making against abstention, so the patron adds no manipulation of his own. The bench's standing compromise still occurs at rate $\lambda > \lambda^\star$ – the rate the rule is designed against, not one it abolishes – but every instance is acquitted, and the capture-induced wrongful-conviction rate is zero. □

The deterrence is in the main ex-post optimal, not merely a threat held in reserve. Because the standing compromise rate is $\lambda > \lambda^\star$, the likelihood ratio at unanimity has inverted below that at $k = n - 1$ (Proposition 1), and acquitting on a unanimous count minimises error in the case at hand as it deters the captor: by Proposition 2, under the maintained $\tau \in (L(n), L(n-1))$, the posterior at $k = n$ lies below τ. The commitment problem is thus milder than the bare threat suggests – but it does not vanish. A sitting bench cannot read the standing rate off the single case before it and may believe *its* unanimity the genuine one; a rule applied panel by panel would invite exactly the discretion a captor leans on. This is a reason for procedure to be *rigid*: a rule fixed in advance and beyond revision in the individual case holds where a bench tempted by an apparently genuine unanimity might not, and rigidity is here a source of the deterrent rather than an administrative convenience. The residual cost of the commitment – acquitting

a genuinely unanimous guilty bench – is of order πp^n, vanishing in the bench size and a price worth paying when wrongful conviction is the graver error. Under strategic voting even this cost lapses in responsive play. By Lemma 1 a responsive free bench reaches a unanimous count only when its decisive judge prefers acquittal, so a unanimous count never warrants conviction – it is capture's manufacture or a free bench's own route to acquittal; acquitting on it is then optimal ex post, not merely ex ante, and the deterrent needs no commitment. Strategic voting does not replace the sincere-voting story but reinforces it. We do not, however, rest this on a probability comparison: whether a strategic bench moves mass onto $k = n - 1$ depends on the equilibrium we have not characterised, and we claim only the outcome-level fact of Lemma 1, not that genuine unanimity becomes rarer in any quantified sense. Commitment is the safeguard for the conservative, sincere-voting case, where a genuine unanimity can occur and the refusal to convict on it must be held against the temptation to defect; the procedure supplies it.

Proposition 5 shows the rule deters capture. We now show more: that among all the rules a court could commit to, it is the one that minimises expected error against a captor who best-responds. Let the designer choose a verdict rule $d : \{0, \dots, n\} \to \{\text{convict}, \text{acquit}\}$ to minimise the expected loss $c_I \Pr(\text{convict}, I) + c_G \Pr(\text{acquit}, G)$, where wrongful conviction is the graver error, $c_I > c_G$, and the cost ratio fixes the threshold of Proposition 2, $\tau = \frac{c_I(1-\pi)}{c_G\,\pi}$. The patron observes d and captures, forcing $k = n$, whenever that is profitable, with a fabricated dissent costing the extra $\delta > B - \kappa$.

Proposition 6 (The rule is the optimal response to capture)**.** *Restricting to rules measurable in the count loses nothing, since a free bench's votes are exchangeable and a compromised bench produces only the all-convict profile. Among all such rules:* (i) *every optimal rule convicts on the interior exactly where $L(k) \geq \tau$, that is on $[k_\tau, n-1]$, because counts $k < n$ are produced only by free benches and so are unaffected by the patron;* (ii) *the only remaining choice is the verdict at $k = n$, and acquitting there is optimal whenever*

$$(1-\pi)\, c_I\, \rho^\star \;>\; c_G\, \pi\, \rho^\star,$$

where $\rho^\star$ is the standing compromise rate a unanimity-honouring rule converts into wrongful convictions – it is the $\lambda > \lambda^\star$ of Proposition 5, positive because a rule that convicts at $k = n$ leaves manufactured consensus profitable ($B > \kappa$) and so neither abolishes the standing rate nor neutralises it. The two sides are the capture losses the rival verdicts incur at $k = n$: convicting wrongly convicts the innocent among the captured benches, $(1-\pi)c_I\rho^\star$; acquitting wrongly acquits the guilty among them, $c_G\pi\rho^\star$. Free benches do not enter the comparison: by Lemma 1 the rule reverses no intended free conviction, and a free bench reaches $k = n$ only when its decisive judge's posterior lies below τ, so acquitting it is pointwise optimal and the convict rule does no better there. The captured comparison favours acquittal exactly when $(1-\pi)c_I > \pi c_G$, that is when $\tau > 1$ – the maintained cost asymmetry, wrongful conviction the graver error. Hence the acquit-at-unanimity rule of Proposition 2 is the loss-minimising commitment against an endogenous

captor whenever the bench is large and wrongful conviction is the graver error.

Proof. (i) For $k < n$ only the free regime contributes, so the conditional loss at each such count is that of a standard binary decision and is minimised pointwise by convicting iff $L(k) \geq \tau$; by Proposition 1 this set is $[k_\tau, n-1]$. These counts and their losses do not depend on the patron's action, so they are common to every rule. *(ii)* At $k = n$ the designer convicts or acquits. If she convicts, blunt capture (cost $\kappa < B$) secures a conviction worth B, so the patron captures; capture forces $k = n$ irrespective of the state, and she then wrongly convicts the innocent among the captured benches, a share $1-\pi$ of them, at cost $(1-\pi)c_I\rho^\star$; the captured guilty she convicts correctly, at no cost. If she acquits, the fabricated-dissent route costs $\kappa + \delta > B$, so the patron adds no conviction and the capture-induced wrongful-conviction rate is zero, though the standing compromise rate $\lambda > \lambda^\star$ persists and is neutralised (Proposition 5). The acquittal does wrongly acquit the guilty among the captured benches – a share π of them, forced to $k = n$ whatever the votes – at cost $c_G\pi\rho^\star$. Free benches do not separate the two rules at $k = n$: by Lemma 1 a free bench reaches $k = n$ only through a decisive judge who prefers acquittal, whose posterior therefore lies below τ, so acquitting it is the pointwise-optimal action and the convict rule does no better there. Acquittal is therefore optimal iff the wrongful convictions it removes on captured benches outweigh the guilty acquittals it incurs on them, $(1-\pi)c_I\rho^\star > c_G\pi\rho^\star$, that is $\tau > 1$, which the maintained cost asymmetry secures; the comparison favours acquittal the more as c_I/c_G rises. □

Capture has a direction. The patron just described captures *toward conviction*; the mirror-image manipulation, a powerful defendant who buys a unanimous acquittal, is a different and, in the setting of a sovereign's courts, a rarer thing. Let conviction-capture occur with intensity λ_c and acquittal-capture with intensity λ_a, and recall that convicting the innocent is the graver error.

Proposition 7 (The asymmetry mirrors the threat)**.** *Suppose conviction-capture is intense enough to invert the top, $\lambda_c > \lambda^\star$. Then for any acquittal-capture intensity λ_a, unanimity is discounted only on the conviction side: the optimal verdict acquits at $k = n$ but does not convict at $k = 0$, since under $\tau > 1$ one has $L(0) < 1 < \tau$ for every λ_a. Only if the error costs reverse, so that the threshold falls below 1, can acquittal-capture invert the bottom. The pro-acquittal tilt strengthens as the capture ratio λ_c/λ_a and the cost of wrongful conviction rise. Were capture to flow toward acquittal past its own threshold and the costs to reverse, the rule would invert. The direction of the rule's suspicion is the direction of capture.*

Proof. Let the free regime carry mass $1-\lambda_c-\lambda_a$, conviction-capture force $k = n$, and acquittal-capture force $k = 0$ (each loads on its own extreme). The interior counts $0 < k < n$ are produced only by the free regime and keep $L(k) = (p/(1-p))^{2k-n}$; the two extremes carry a contamination

term in their own capture intensity,

$$L(n) = \frac{(1-\lambda_c-\lambda_a)\,p^n + \lambda_c}{(1-\lambda_c-\lambda_a)(1-p)^n + \lambda_c}, \qquad L(0) = \frac{(1-\lambda_c-\lambda_a)(1-p)^n + \lambda_a}{(1-\lambda_c-\lambda_a)\,p^n + \lambda_a}.$$

By the argument of Proposition 1, $L(n)$ is strictly decreasing in λ_c and falls below $L(n-1)$ once $\lambda_c > \lambda^\star$, and hence below τ (the second step using the maintained $\tau < L(n-1)$ of Proposition 2; inversion alone would not place the count below the threshold); the optimal conviction set then excludes $k = n$. The acquittal end behaves differently. Although $L(0)$ rises in λ_a, it does so only toward 1 from below – it begins at $((1-p)/p)^n < 1$ and the contamination cannot carry it past 1 – so $L(0) < 1 < \tau$ throughout, and $k = 0$ never warrants conviction under the maintained costs, whatever λ_a; the acquittal-side threshold $\lambda_a^\star$ at which $L(0)$ would clear the relevant threshold is reached only once the costs reverse so that that threshold falls below 1. The asymmetry is thus not an artefact of assuming $\lambda_c > \lambda_a$: even with the two intensities equal, the convict-consensus can be discounted to the point of flipping the verdict, while the acquit-consensus cannot, because flipping it would mean convicting on a near-unanimous acquittal, which the high threshold τ forbids. Given $\lambda_c > \lambda^\star$, the discount grows in λ_c/λ_a and in the cost of wrongful conviction; reversing the costs and pushing λ_a past $\lambda_a^\star$ reverses the rule. □

The procedure's whole pro-acquittal apparatus – a majority of one to acquit but two to convict, the opening for acquittal, the bar on a unanimous conviction – reads here as the institutional trace of a single asymmetry: in a polity whose courts are pressed toward conviction, consensus is dangerous only when it condemns. The device is the one Solomon uses and the implementation literature studies (Glazer and Ma, 1989; Moore, 1992): an off-equilibrium response that makes manipulation unprofitable, so that on the path it need never be used.[2]

7.1 When capture loads on consensus

That capture manufactures *unanimity* – rather than a managed majority with a tolerated dissent – is the premise on which the deterrence rests, and it is not self-evident. Captured courts often stage disagreement. A packed bench may seat a minority precisely to project deliberation; a backsliding regime that controls a court's direction packs it for a reliable majority, not for an n-for-n tally, and leaves the independents to dissent on the record (Poland and Hungary are the contemporary instances; the autocracy literature reads manufactured *margins*, not manufactured consensus, as the signature of control – Magaloni, 2006; Simpser, 2013). If a captor can cheaply seat a four-to-one, the assumption that consensus is the cheap manipulation and credible dissent the dear one is exactly wrong. Two features of the procedure answer the objection, and together

[2]The analogy is to the off-path-threat logic, not to the environment: those are full-information implementation results, whereas the court's problem is one of incomplete information about guilt. What carries over is only that a credible response to a deviation – here, refusing to convict on the count capture forces – can make the deviation unprofitable and so keep it off the path.

they mark the scope of the rule.

The first is the threat the rule is built against. Capital procedure is already pro-acquittal: a majority of one acquits where two are needed to convict, and a judge may move toward acquittal but not away from it (Sanhedrin 4:1). A single un-suborned judge inclined to acquit therefore defeats a conviction. A patron who wants *this* defendant condemned with certainty cannot leave a pivotal vote outside his control; to be sure of the verdict he must drive the whole bench. Capture loads on unanimity here not as a law of captured courts in general but because, under an acquittal-protective margin, certainty of conviction *requires* consensus. Where a captor seeks only to steer a court's output, a bare majority governs and dissent survives on the record – and there the unanimity rule is the wrong instrument and does not bind. The rule answers the first threat, the guaranteed conviction of a marked defendant, and not the second, the directional control of a docket.

The second is that the cost of a fabricated dissent is not assumed but built. The procedure records each judge's stated reasons – two scribes write down the words of those who would acquit and of those who would convict (Sanhedrin 4:3) – preserves the minority view as a thing a later court may rely upon (Eduyot 1:5), and reconsiders the case overnight. A manufactured dissent is then not a silent vote but a reasoned, recorded, examinable opinion whose author is an exposed co-conspirator with a night in which to defect. Staging credible disagreement is dear precisely where dissent must carry its reasons and survive scrutiny; the gap δ between manufacturing consensus and manufacturing a credible dissent is the procedure's own work. Where dissent is decorative – unreasoned, unrecorded, unexamined – a minority is cheap to seat, δ is small, and the rule would not bite. The acquit-at-unanimity rule and the recorded-reasoned-dissent requirement are thus complements: neither deters capture without the other.

The managed-dissent courts are then not counterexamples but cases outside the rule's scope. A bench that seats a tolerated minority for the appearance of deliberation, or that is packed for a governing majority, satisfies neither condition: it faces no acquittal-protective margin that forces the captor to consensus, and it records dissent in no form that makes a fabricated one costly. The model accordingly predicts where the rule should and should not appear – a refusal to convict on unanimity belongs to procedures that also protect acquittal and compel reasoned, recorded dissent, and is pointless where a bare majority decides or where dissent is for show. That the Sanhedrin pairs the unanimity rule with exactly these features – the acquittal-protective margin, the recorded reasoned vote, the overnight reconsideration – is the evidence that it is the design the theory describes.

The argument so far has taken the compromised bench to return exact unanimity, and one should ask whether the rule survives a captor who blurs it. A bench that can stage a managed dissent puts mass not at $k = n$ but just below it; the lone dissent is then no longer impossible under capture, the count $k = n - 1$ is itself contaminated, and the authentication it once carried weakens. We show the keystone is not a knife-edge but a margin. Let the compromised regime

return $k = n$ with probability $1 - \varepsilon$ and $k = n - 1$ with probability ε, state-independently – the sharpest such perturbation, a single staged dissent.

Proposition 8 (Robustness to staged dissent)**.** *There is an $\bar{\varepsilon} > 0$ such that for every $\varepsilon < \bar{\varepsilon}$ the optimal verdict is unchanged: it acquits at $k = n$ and convicts on the interval $[k_\tau, n-1]$. The verdict at each count varies continuously in ε and converges to that of Proposition 2 as $\varepsilon \to 0$. The margin is*

$$\bar{\varepsilon} = \min\left\{ 1 - \frac{(1-\lambda)\big(p^{\,n} - \tau(1-p)^{\,n}\big)}{\lambda(\tau - 1)},\quad \frac{(1-\lambda)\,n\big[p^{\,n-1}(1-p) - \tau(1-p)^{\,n-1}p\big]}{\lambda(\tau-1)} \right\},$$

both terms positive under the maintained conditions $\lambda > \lambda^\star$ and $\tau \in \big(L(n), L(n-1)\big)$.

Proof. Under the perturbation $\Pr(n \mid \omega) = (1-\lambda)\rho_\omega^n + \lambda(1-\varepsilon)$ and $\Pr(n-1 \mid \omega) = (1-\lambda)\,n\,\rho_\omega^{n-1}(1-\rho_\omega) + \lambda\varepsilon$, with $\rho_G = p$, $\rho_I = 1-p$; counts $k < n-1$ are untouched and keep $L(k) = (p/(1-p))^{2k-n}$. Both $L(n)$ and $L(n-1)$ are ratios of functions affine in ε, hence continuous and monotone in it. At the top the effective atom $\lambda(1-\varepsilon)$ shrinks as ε rises, so $L(n)$ increases from its Proposition 1 value toward the clean $(p/(1-p))^n$; it reaches τ when $\lambda(1-\varepsilon) = (1-\lambda)\big(p^{\,n} - \tau(1-p)^{\,n}\big)/(\tau-1)$, that is at the first bound, positive precisely because $L(n) < \tau$ at $\varepsilon = 0$. At $k = n-1$ the atom $\lambda\varepsilon$ grows, dragging $L(n-1)$ down from $(p/(1-p))^{n-2}$ toward 1; it reaches τ at the second bound, positive because $L(n-1) = (p/(1-p))^{n-2} > \tau$ and $\tau > 1$. For ε below both, $L(n) < \tau < L(n-1)$ as in Proposition 2, the interior is unmoved, and the conviction set remains $[k_\tau, n-1]$. Continuity of the two ratios gives convergence to the baseline as $\varepsilon \to 0$. □

So the rule's structure moves continuously, not discontinuously, as the captor's consensus loosens: a staged dissent does contaminate the count that convicts, and the lone dissent authenticates less surely the more readily it can be faked, but so long as the staged dissent is rare relative to the standing pull toward consensus – so long as $\varepsilon < \bar{\varepsilon}$ – the rule that acquits on unanimity and convicts on a single dissent remains the optimal one. The same holds for a general compromised distribution with sub-unanimity mass, once $\bar{\varepsilon}$ is redefined as the minimum, over $k = n$ and every convicted interior count $k \in [k_\tau, n-1]$, of the slack to τ at that count divided by the largest state-independent atom it can receive; the rule is unchanged for total contaminating mass below this (tighter) $\bar{\varepsilon}$. The single staged dissent above is the case in which only $k = n$ and $k = n-1$ are affected. And the margin is exactly what the recorded-dissent machinery of this subsection secures: making a fabricated dissent dear is what holds ε small.

7.2 Robustness to capture

A captor seeking the count $k = n-1$ that the rule still honours has two routes. He may leave one judge uncaptured and hope the verdict falls there, or capture the whole bench and fabricate

a dissent. The first route turns on what the free judge infers, because the captor chooses *which* judge to leave free and the voting order is fixed.

Under partial capture the bench is neither wholly free nor wholly compromised: a faction holds all but one seat, the $n-1$ suborned judges are instructed to convict, and one judge is free. This is the third regime of the type space of Section 2, standing at rate $\mu > 0$ alongside the free and the fully compromised benches. Like λ, the rate μ is exogenous background risk rather than a quantity the captor tunes in equilibrium; the proposition shows that against it no *added* attack pays, so there is no fixed point in μ to solve. The lone free judge does not know he faces it; conditioning on his own freedom – which rules out the fully compromised regime, in which he too would have been instructed – he weighs the free regime against the partial one and derives, rather than assumes, what an apparent consensus before him portends. A faction holding $n-1$ seats manufactures exactly that near-consensus, and at a standing rate $\mu > 0$ the possibility is on the equilibrium path, not a conjecture about a deviation. Whether the inference bites turns on where the order places him.

Proposition 9 (Partial capture fails)**.** *Suppose the captor suborns $n-1$ judges and leaves one free, voting is junior-first (Sanhedrin 4:2), the partially compromised regime stands at rate $\mu > \mu^\star$, the cost of suborning a judge rises in his seniority, and the bench is large enough that $(1-p)B \le (n-1)c$ and $(n-1)c + \Delta \ge pB$. Then no partial attack is* profitable. *Leaving the* senior *free is the cheap route: moving last, he sees an apparent near-unanimity and discounts it exactly as the rule discounts unanimity, so the contaminated prefix no longer carries his vote and he convicts the innocent only on his own erroneous signal, at rate at most $1-p$; the bench prices this residual out whenever $(1-p)B \le (n-1)c$. Leaving an early* junior *free raises the success rate to p – voting before any consensus is on the record, he has nothing to discount and dissents on an innocent signal, leaving $k = n-1$ – but it must suborn the senior in his place, and the seniority premium Δ together with the $n-1$ per-judge costs price it out whenever $(n-1)c + \Delta \ge pB$.*

Proof. The lone free judge at the penultimate node sees $n-1$ convicting votes. As at $k = n$ in Proposition 1, the prefix is contaminated: a free guilty bench produces it with probability proportional to $p^{\,n-1}$, a free innocent bench with probability proportional to $(1-p)^{\,n-1}$, and a partially compromised bench – which, conditioning on his own freedom, is the only compromise he need weigh – produces it with certainty, independently of the state. The prefix likelihood ratio

$$L_{n-1}^{\text{seq}} = \frac{(1-\lambda-\mu)\,p^{\,n-1} + \mu}{(1-\lambda-\mu)(1-p)^{\,n-1} + \mu}$$

starts above τ at $\mu = 0$, since $(p/(1-p))^{\,n-1} > L(n-1) > \tau$, is strictly decreasing in μ by the affine-ratio argument of Proposition 1, and so crosses τ at a unique $\mu^\star$, the sequential analogue of that proposition's $\lambda^\star$; for $\mu > \mu^\star$ it lies below τ. Because the regime it discounts stands at a positive rate, the inference is on the equilibrium path, not a belief about a deviation. The

senior's own signal moves the posterior by at most the factor $p/(1-p)$: against the discredited prefix it tips him to conviction only when it convicts, so he delivers $k = n-1$ on his own erroneous signal alone, with probability at most $1-p$ (and, when $\tau > p/(1-p)$, vanishing as $\mu \to 1-\lambda$). The senior-free route thus convicts the innocent at rate at most $1-p$ for the price of the $n-1$ juniors, a payoff of at most $(1-p)B-(n-1)c$, negative when $(1-p)B \leq (n-1)c$. A free junior moving first sees no prefix to discount: he votes his signal and leaves $k = n-1$ at rate p. But that route leaves a junior, not the senior, free, hence suborns the senior at premium Δ; its payoff is $pB-(n-1)c-\Delta$, negative when $(n-1)c+\Delta \geq pB$. Full capture with a fabricated dissent returns $B-\kappa-\delta < 0$. Abstention dominates, and no partial attack is profitable. □

The suspicion is the rule's own logic turned on the bench, and it is consistent rather than conjectural. Because partial capture stands at a positive rate $\mu > \mu^\star$, the lone free judge's discount of an apparent consensus is the very inference the designer makes at $k = n$ in Proposition 1 – self-discrediting consensus – now read off a regime that is on the path; it is not a belief about a deviation, and no consistency refinement disturbs it. A judge schooled to distrust manufactured consensus will not ratify it on the strength of the prefix alone, so the cheap and natural partial attack – leave the senior free, suborn the juniors – can convict only on his own error, at rate at most $1-p$, and the bench size then prices that residual out. The information structure collapses the attack's success from the near-certainty a credulous senior would hand it to the senior's own error rate; the per-judge cost closes what remains.

The seniority premium Δ closes the one partial route the suspicion does not touch: the junior-free attack, in which the free judge votes too early to have anything to discount. A reader tallying assumptions should see the wedges plainly – the gap δ between manufacturing consensus and manufacturing a credible dissent, which closes the full-capture route; the seniority premium Δ, which closes the junior-free route; and the per-judge cost c, which closes the residual of the senior-free route once the suspicion has capped its success at $1-p$. The connection to the graded benches is already visible: the junior-free route is priced out by $(n-1)c+\Delta \geq pB$ and the senior-free route by $(n-1)c \geq (1-p)B$, both thresholds easier to clear the larger the bench. Abler judges – a higher p – cut the two routes apart: a higher p lowers the senior-free cap $(1-p)B$ and so deters that route, but raises the junior-free prize pB and so strains its deterrent. And where $\tau > p/(1-p)$ the suspicion bites harder still, lowering the senior-free cap below $1-p$ toward zero as μ grows and the consensus it must discredit lengthens. The sequential, public order is essential: it is what lets the senior, moving last, see the apparent consensus and refuse to ratify it, and what denies the captor a free seat early in the order without paying for the senior.

The remaining route is to capture all n and instruct one judge to dissent, manufacturing $k = n-1$ directly. This is the only place the cost δ of a fabricated dissent is required, and the procedure makes it positive by design, as the previous subsection set out: the capital bench records each judge's stated reasons (Sanhedrin 4:3) and preserves the minority view (*Eduyot*

1:5), so a fabricated dissent is not a silent vote but a reasoned minority opinion that must survive examination, and the planted dissenter is a co-conspirator who may defect. The rule deters when $B < \kappa + \delta$: blunt capture acquits, and disguised capture is too costly to disguise.

7.3 Commitment and the foundations of the deterrent

With the standing rate $\lambda > \lambda^\star$ the rule is, as the last section showed, in the main ex-post optimal: a unanimous count warrants acquittal in the case at hand. What commitment must guard is the knife-edge – the bench that, unable to read the standing rate off its own case, believes its unanimity the genuine one and is tempted to convict. The threat must remain credible there, beyond the reach of the individual panel, and the procedure supplies three bindings.

First, the verdict is a *mechanical function of the count*, not a judgement the panel forms about whether a given unanimity is suspect. Discretion is removed at the point of application, and the map from votes to verdict is fixed by the tradition, not by the sitting bench. This is commitment through rules rather than discretion (Kydland and Prescott, 1977).

Second, the rule is public and codified, and may be overturned only by a court *greater in wisdom and number* (*Eduyot* 1:5) – a deliberately high cost of revision, so that no single bench can set the rule aside in the case before it.

Third, the commitment is sustained by repetition, and this we can state exactly. Cases arrive over time; a court that convicted on a unanimous bench would teach every future patron that the threat is empty, and the deterrent would unravel across all subsequent cases. Under the grim-trigger profile – acquit at unanimity always, revert to convicting on it once the rule is broken – acquit-at-unanimity is subgame-perfect whenever the court is patient enough, $\beta \geq g/(g+D)$, with g the one-time gain from convicting a genuinely-unanimous guilty bench and D the per-period deterrence value the rule preserves (Proposition 12, proved in the appendix). The threshold $\beta^\star = g/(g+D)$ does not depend on how often the tempting history arises; rarity does not relax it. What rarity buys is that the constraint binds seldom – the tempting history has probability of order πp^n, vanishing in the bench size, and under strategic voting never occurs at all (Lemma 1) – so the same large benches that price out capture are those on which the commitment is almost never tested. The shadow of future cases, not a one-shot promise, holds the court to its rule.

7.4 Capture with and without consensus

The theory is built on a historical court, but its content is general, and it is worth asking what it says about a captured court of the present. Take the clearest contemporary instance, Poland's Constitutional Tribunal after 2015. A governing party that wished to control constitutional review packed the Tribunal – contesting appointments, seating loyalists, declining to recognise judges lawfully named before it – until the bench returned the rulings the government wanted

(Nalepa, 2022; Chiopris et al., 2025). This is capture, and it is real. But it is not the capture this paper's rule answers, and the difference is exactly the scope condition.

Poland's was capture for *direction*: control of which laws survive review, exercised through a reliable majority on a standing court whose business is a stream of cases, not the conviction of a single marked defendant. A governing majority suffices for it; the captor need not drive the bench to unanimity, and the Tribunal's dissenters went on dissenting, on the record. By the model this is the out-of-scope regime – no acquittal-protective margin forces the captor to consensus, and a court of constitutional review records dissent as a matter of course, so manufacturing it is not dear. The acquit-at-unanimity rule would do nothing here: it audits a signature this capture has no reason to produce.

The other branch is visible in the show trial. When a regime wants a *particular* adversary condemned – an Old Bolshevik in the Moscow trials, a purged officer before a military collegium – the threat is exactly the one the Sanhedrin's procedure is built against: a guaranteed verdict against a marked defendant, which a single honest vote to acquit would spoil. There capture does drive consensus. The tribunals returned their convictions without dissent, the defence hollow, the verdict settled before the hearing; manufactured unanimity is the genre's signature. But they manufacture it cheaply, because the regime that runs them has first stripped away everything that would make a dissent costly to suppress – the real defence, the recorded reasons, the independent record. That is what the comparison fixes: the Sanhedrin's rule is not a description of how captured courts behave but a design *against* them, raised by a court-builder who anticipates the show trial and makes its manufactured consensus self-defeating rather than decisive.

Taken together the two courts fix the theory's content. The model sorts anti-capture instruments by the structure of the threat. Where the danger is a guaranteed verdict against a particular person, under a procedure that one vote can defeat and that compels reasoned, recorded dissent, the designer's instrument is to discredit consensus – the Sanhedrin's rule. Where the danger is directional control of a docket by majority, the consensus signature is absent and the instrument must be otherwise: random assignment of panels, recusal, super-majority confirmation of judges, fixed terms, the publication of reasons. The prediction that travels is therefore comparative and testable: rules that refuse to convict on unanimity should appear in high-stakes, acquittal-protective, single-defendant procedures that record reasoned dissent, and should be absent from courts of review where a majority governs – while the anti-capture apparatus of the latter takes the different form its different threat calls for. A consensus-discrediting rule grafted onto a constitutional court would be a category error; its absence there is the theory confirmed, not contradicted.

8 The size of the bench

The Mishnah grades the court by the gravity of the matter: three judges for money, twenty-three for a capital charge, and seventy-one for the gravest causes of the nation (Sanhedrin 1:1–1:6). The grades are themselves read from Scripture (Sanhedrin 2a): the capital twenty-three from the verse's two congregations, one that judges and one that delivers – *ve-shafetu ha-edah… ve-hitzilu ha-edah* (Numbers 35:24–25), each an *edah* of ten, with three more added so that a convicting majority can stand against an acquitting one; the great court of seventy-one from the seventy elders gathered to Moses, *esfah li shiv'im ish* (Numbers 11:16), with Moses over them. The capture theory explains why the grading takes the direction it does. Take the captor's most effective *partial* attack (Proposition 9). Of its two routes, the senior-free attack corrupts the $n-1$ *junior* judges at cost $(n-1)c$ and convicts the innocent with probability at most $1-p$ – the free senior's error rate being the ceiling on the suspicious judge's compliance – while the junior-free attack convicts at the higher rate p but must buy the senior, at the seniority premium Δ. The captor takes whichever pays more, so the bench's worst case is $\max\{(1-p)B-(n-1)c,\ pB-(n-1)c-\Delta\}$; senior-free is the cheaper route only when $\Delta > (2p-1)B$, and otherwise the captor prefers the abler junior-free attack. Here c is the junior cost throughout, and corrupting the whole bench costs at least $nc+\delta$ once a dissent must be fabricated (the senior premium only raises that cost, reinforcing deterrence); that full route attacks only if $B > nc+\delta$.

Proposition 10 (The bench is sized to price out capture)**.** *A bench deters capture once*

$$n \ \geq\ n^*(B,p,c,\Delta,\delta) = \left\lceil \max\left\{ 1+\frac{\max\{(1-p)B,\ pB-\Delta\}}{c},\ \frac{B-\delta}{c} \right\} \right\rceil,$$

a sufficient threshold increasing in the captor's stake B. Since graver charges raise the value B of securing a conviction, they require larger benches: the grading $3 < 23 < 71$ is the bench scaling with the gravity of the matter.

Proof. By Proposition 9 the lone free judge's consistent suspicion caps the cheap senior-free attack at the senior's error rate $1-p$, so it nets at most $(1-p)B-(n-1)c$, non-positive once $n \geq 1+(1-p)B/c$. The junior-free route nets $pB-(n-1)c-\Delta$ – the conviction rate is the free junior's ability p, and the route must pay the seniority premium Δ – non-positive once $n \geq 1+(pB-\Delta)/c$; the captor takes whichever partial route pays more, so both are deterred once $n \geq 1+\max\{(1-p)B,\, pB-\Delta\}/c$. The fabricated-dissent attack corrupts all n judges and stages a dissent, costing at least $nc+\delta$ for a conviction worth B, so it nets at most $B-nc-\delta$, non-positive once $n \geq (B-\delta)/c$. All three routes are unprofitable once $n \geq \max\{1+\max\{(1-p)B,\, pB-\Delta\}/c,\ (B-\delta)/c\}$; since each cost figure is a lower bound on the captor's outlay, this n^* is a sufficient deterring size rather than the exact minimum. Both bounds increase in B. □

Two features reinforce the result. First, the rule and the large bench are complements. The cost of committing to acquit at unanimity is the occasional loss of a genuine unanimous conviction, of order πp^n, which *vanishes* as n grows: on a bench of twenty-three a free-bench unanimity is almost never seen, so the deterrent is almost costless to maintain. Second, the same enlargement that prices out capture aggregates more independent signals and lowers the error rate, which matters most exactly where error is gravest. Deterrence and accuracy call for size together.

Remark 5 (Interior benches under uncertain stakes). If the captor's stake B is private, drawn from a distribution G, a bench of size n deters all captors with B below a threshold increasing in n, leaving a residual capture probability that falls in n. Trading this against the cost of convening and vetting judges yields an interior optimal bench that rises with the gravity of the cause – the grading as a smooth comparative static rather than a knife-edge.

9 The order of voting

A capital bench states its opinions "from the side" – the junior judges first – so that they not be swayed by the greatest among them (Sanhedrin 36a). We read the order as the device that manufactures the independence on which the rest of the procedure relies. Let judges differ in seniority, and let the bench, with probability $\gamma > 0$, fall into *deference*: a common cascade in which every judge who has already heard a senior vote copies him in place of his own signal. The deference we posit is a reduced form for the reputational herding of Ottaviani and Sørensen (2001), in which experts anxious to appear well informed echo those who have spoken before them; we take γ as a primitive rather than deriving it, and the aggregative content of what follows – that junior-first preserves independent signals – is theirs, not ours. What we add is the capture overlay: junior-first is also what denies a captor the deference shortcut, and so protects the bench-size deterrent. Under *senior-first* voting a deference cascade sweeps the whole bench from the first senior's vote; under *junior-first* voting the juniors commit their signals before any senior speaks, the cascade has no senior to form on, and every vote is informative.

Proposition 11 (Junior-first and the unanimity rule are complements)**.** *Junior-first voting yields independent votes, so an honest bench convicts unanimously with probability of order p^n, vanishing in n. Senior-first voting admits the deference cascade, so honest unanimity occurs with probability at least $\gamma \Pr(\textit{senior convicts})$, a positive constant independent of n. Hence junior-first both minimises the commitment cost $\sim \pi \Pr(\textit{honest unanimity})$ of the acquit-at-unanimity rule and maximises its power to detect capture: honest unanimity becomes rare while captured unanimity does not. The order manufactures the independence the unanimity rule audits.*

Proof. Under junior-first voting a junior moves before any senior, so no cascade can form and every vote reflects an independent signal; an honest bench is then unanimous only if all n signals

agree, with probability p^n under G and $(1-p)^n$ under I, of order $p^n \to 0$. Under senior-first voting a deference cascade forms with probability γ, and conditional on it every judge copies the first senior, so the bench is unanimous whenever the senior votes to convict; honest unanimity therefore occurs with probability at least $\gamma \Pr(\text{senior convicts})$, which does not depend on n. (The independent-copying alternative, in which each junior defers separately with probability γ, gives honest unanimity $[\gamma+(1-\gamma)p]^{n-1} \to 0$ and would not separate the orders; it is the *common* cascade, a single correlated failure of independence, that the junior-first order forecloses.) The commitment cost $\pi \Pr(\text{honest unanimity})$ is thus minimised under junior-first, while captured unanimity occurs with probability one regardless of order; the likelihood ratio of captured to honest unanimity, the rule's detection power, is therefore maximised under junior-first. □

Remark 6 (Junior-first protects the bench-size deterrent). Under deference a captor need only corrupt the most senior judge, whom the rest copy, so the cost of swaying the bench collapses to $c+\Delta$ regardless of n and the grading of Proposition 10 loses its bite. Junior-first removes the shortcut: with the juniors committed in advance, capturing the senior changes nothing, and the captor must corrupt judges one by one, restoring the $(n-1)c$ cost on which deterrence rests. The same public sequence is what lets the lone free judge of Proposition 9 see the apparent consensus and refuse to ratify it. The order thus does triple duty – independence, the free judge's sight of the consensus, and the integrity of the bench-size deterrent.

10 Relation to the literature

The Condorcet tradition and its strategic descendants (Austen-Smith and Banks, 1996; Feddersen and Pesendorfer, 1998) assume conditional independence, under which the count is informative-monotone and optimal rules are thresholds; Feddersen and Pesendorfer (1998) show that the unanimity rule can be especially poor once voting is strategic. Duggan and Martinelli (2001) carry the analysis to a continuum of signals and show the outcome turns on the likelihood ratio: where it is bounded, unanimity can leave the probability of error bounded away from zero as the jury grows, while where it is unbounded, unanimity can still aggregate information asymptotically. Our point is adjacent and distinct: when the designer is uncertain whether independence holds at all, the optimal use of the vote is non-monotone, and a unanimous verdict is discounted rather than trusted. The primitive is not a richer signal structure but a doubt about the procedure that generated the signals.

A recent and prominent case for abandoning unanimity is made by Bouton et al. (2018), who show that majority rules with veto power Pareto-dominate the unanimous rules and are ex ante efficient across a broad class of environments. Their verdict and ours run the same way – a unanimous standard is not to be trusted – by opposite routes. Theirs is a fully specified positive model: voters with one-sided preferences play an equilibrium under a fixed rule, and the rules are ranked by the welfare their equilibria deliver; the veto earns its place by shielding

the decision against a partisan voter who would exploit a unanimity rule. Ours is a design: we hold the asymmetry of error in the objective and ask what mapping from counts to verdicts an optimizing court should adopt, reading the observed rule off the answer; the acquittal at unanimity earns its place by shielding the *verdict* against a bench that may have been captured. The mechanisms differ accordingly. In their account the count stays informative-monotone, and unanimity fails through the strategic incentives it sets and the welfare it forgoes; in ours the count inverts at the top, so a unanimous conviction is *less* probative of guilt than a lone dissent, and conviction is withheld because the evidence has turned. And the objects differ: their target is the unanimity *rule*, the requirement that all concur to act, while ours is the unanimous *verdict* as an event, on a bench that convicts by a majority of two. We do not impose unanimity and find it wanting; we observe it and read it as exculpatory. The mechanism we exploit is not theirs but the one-sided, state-independent limit of the correlated-votes tradition we turn to next.

The dependence we introduce ties the argument to the literature on *correlated* votes in the Condorcet setting, where a common influence across jurors blunts aggregation and can overturn the asymptotic jury theorem (Ladha, 1992): capture is the sharp limit of such correlation, a common cause that moves the whole bench at once. Ladha is the nearest antecedent, and it marks what is and is not new here. The non-monotone reading of the count is not new; it is the degenerate limit of his correlated votes. What is new is that the correlation is *one-sided* and *state-independent*, loaded on the convict-consensus a captor manufactures whether or not the defendant is guilty – which is what inverts the posterior at the top, as the symmetric correlation Ladha studies does not – and that the rule answers it by *design* – it is built to deter the very dependence it fears. That design turn places the argument as much in the theory of collusion and manipulation as in the theory of aggregation: the deterrence results below read the unanimity rule as a device that makes suborning the bench unprofitable, in the manner of the mechanism-design treatment of collusion in organisations (Tirole, 1986; Laffont and Martimort, 1997), and the contribution relative to the jury literature is not the statistics of the contaminated count but the use of a verdict rule to price manipulation out. And where Coughlan (2000) defends the unanimity rule by letting jurors communicate before voting – so that, when their thresholds are sufficiently aligned, sincere voting is an equilibrium and unanimity minimises error – we defend a rule that *withholds* conviction from unanimity for a different reason: not to coordinate honest jurors but to disarm a bench that may not be honest. The two defences are complementary, the one addressing strategic voting under aligned preferences and the other procedural capture.

That the right rule turns on the quality of the information is a theme of optimal committee design with *endogenous* information: Persico (2004) shows that the voting rule feeds back on the jurors' incentive to acquire costly signals, so that a demanding supermajority, even unanimity, is optimal only when signals are accurate enough to be worth gathering under it. Our graded benches share the comparative static – scaling the court to the gravity, and so the stakes, of the cause – though our channel is the deterrence of capture rather than the provision of acquisition

incentives. And once pre-vote deliberation is allowed, a large class of voting rules induce the same set of equilibrium outcomes, the unanimity rules being the known exception (Gerardi and Yariv, 2007); it is exactly at unanimity that our procedure parts company with the monotone rules, there withholding conviction rather than conferring it.

The order of voting has a literature of its own, and its formal home is Ottaviani and Sørensen (2001), who model a debate among reputation-minded experts and ask in what order they should speak. Their answer is ours in spirit – letting the junior speak first denies him the chance to defer, so his signal enters the record undistorted – and they too read the rule off the capital bench, citing the Mishnah that opens the count "from the side." Two things separate the arguments. Their herding is reputational, a wish to appear well informed; the dependence we fear is capture, a bench that may not be honest at all, and junior-first earns its place here not by improving aggregation but by denying a captor the deference shortcut, and so protecting the bench-size deterrent. And where they find the anti-seniority rule *not always* optimal for aggregation – a more expert member speaking later may herd on a weaker earlier one – the Talmud fixes junior-first without qualification; the capture rationale supplies the reason a rule of ambiguous aggregative value is nonetheless made unconditional. The two readings meet at the same Mishnah from opposite sides: they begin from the bench to reach a theorem of debate, while we take the bench as the object and ask why both its rules are there.

A A repeated-game foundation for the deterrent

Under sincere voting the rule is in the main ex-post optimal (Section 7), but a court that has seen no capture for a long while may come to believe a unanimous bench before it genuine – unable to read the standing rate off its own case – and be tempted ex post to convict it. This is the knife-edge the commitment subsection identified. The third binding holds that repetition sustains the rule against that temptation. We state and prove it.

Cases arrive in discrete periods, one per period, with common discount factor $\beta \in (0,1)$. In each period a patron may capture the bench, forcing $k = n$, if he expects conviction there; the court, unable to bind its hand mechanically, chooses afresh at each unanimous count whether to acquit or convict, and patrons observe the court's record. Let $g > 0$ be the one-period gain from convicting a genuinely-unanimous guilty bench rather than acquitting it – the largest the temptation can be, since at a unanimous count the court cannot tell a genuine bench from a manufactured one, and $\lambda > \lambda^\star$ makes the manufactured the likelier, so the true expected deviation gain is smaller and the bound $\beta^\star$ correspondingly slacker. Let $D > 0$ be the per-period deterrence value: the wrongful convictions the standing rule prevents, $(1-\pi)c_I\rho^\star$, net of the correct convictions it forgoes by acquitting unanimous benches – the guilty among the captured, $c_G\pi\rho^\star$, and, under sincere voting, the genuinely-unanimous free guilty, $c_G\pi p^n(1-\rho^\star)$. Thus $D = (1-\pi)c_I\rho^\star - c_G\pi\rho^\star - c_G\pi p^n(1-\rho^\star)$, with $\rho^\star$ the standing compromise rate of

Proposition 5; the rule is worth sustaining only where $D > 0$, prevented wrongful convictions outweighing forgone correct ones. Under strategic voting the free term vanishes and $D > 0$ reduces to $\tau > 1$, the criterion of Proposition 6.

Proposition 12 (The deterrent is self-enforcing)**.** *Under the grim-trigger profile – the court acquits at unanimity in every period and reverts permanently to the no-commitment play (convict at unanimity, so manufactured consensus convicts again and the wrongful-conviction rate jumps back to $(1-\pi)\rho^\star$) the first time it convicts on a unanimous bench – acquittal at unanimity is a subgame-perfect equilibrium if and only if*

$$\beta \;\geq\; \beta^\star \;=\; \frac{g}{g+D}.$$

The tempting history – a genuinely-unanimous free bench the court would rather convict – occurs under sincere voting with per-period probability of order πp^n, vanishing in n; under strategic voting it is off the path, since a free bench reaches $k = n$ only through a decisive judge who prefers acquittal (Lemma 1), and the rule is then self-enforcing for every β.

Proof. By the one-shot deviation principle it suffices to check each history. At the only history where the prescribed acquittal is not statically optimal – a genuinely-unanimous, and hence almost surely guilty, bench under sincere voting – acquitting forgoes the immediate gain g but preserves the deterrent, worth D in every future period, while convicting takes g now and, being observed, triggers reversion that forfeits D thereafter. Acquittal is optimal there iff $g \leq \frac{\beta}{1-\beta}D$, that is $\beta \geq g/(g+D)$. The verdict record is public, so reversion fires on observed play and the off-path beliefs supporting it are trivial. At every other history the prescribed play is statically optimal: on the interior the increasing threshold verdict, and at $k = n$ – where a responsive free bench arrives only through a decisive judge who prefers acquittal – the prescribed acquittal is itself statically optimal, so no further one-shot deviation is profitable, and the condition is necessary and sufficient. Under strategic voting a free bench that would convict stops at $k = n-1$ rather than complete the unanimity (Lemma 1), so no guilty-leaning bench reaches $k = n$; the tempting history is off the path, acquittal at unanimity is ex post optimal, and the rule holds for any β. □

References

Aumann, R. and M. Maschler (1985). "Game Theoretic Analysis of a Bankruptcy Problem from the Talmud." *Journal of Economic Theory* 36(2), 195–213.

Austen-Smith, D. and J. Banks (1996). "Information Aggregation, Rationality, and the Condorcet Jury Theorem." *American Political Science Review* 90(1), 34–45.

Banerjee, A. V. (1992). "A Simple Model of Herd Behavior." *Quarterly Journal of Economics* 107(3), 797–817.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992). "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy* 100(5), 992–1026.

Bouton, L., A. Llorente-Saguer, and F. Malherbe (2018). "Get Rid of Unanimity Rule: The Superiority of Majority Rules with Veto Power." *Journal of Political Economy* 126(1), 107–149.

Chiopris, C., M. Nalepa, and G. Vanberg (2025). "A Wolf in Sheep's Clothing: Citizen Uncertainty and Democratic Backsliding." *Journal of Politics* 87(4), 1272–1287.

Condorcet, M. de (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.*

Coughlan, P. J. (2000). "In Defense of Unanimous Jury Verdicts: Mistrials, Communication, and Strategic Voting." *American Political Science Review* 94(2), 375–393.

Dekel, E. and M. Piccione (2000). "Sequential Voting Procedures in Symmetric Binary Elections." *Journal of Political Economy* 108(1), 34–55.

Duggan, J. and C. Martinelli (2001). "A Bayesian Model of Voting in Juries." *Games and Economic Behavior* 37(2), 259–294.

Feddersen, T. and W. Pesendorfer (1998). "Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting." *American Political Science Review* 92(1), 23–35.

Gandhi, J. (2008). *Political Institutions under Dictatorship.* Cambridge University Press.

Gerardi, D. and L. Yariv (2007). "Deliberative Voting." *Journal of Economic Theory* 134(1), 317–338.

Glazer, J. and C.-T. A. Ma (1989). "Efficient Allocation of a 'Prize' – King Solomon's Dilemma." *Games and Economic Behavior* 1(3), 222–233.

Groseclose, T. and J. M. Snyder (1996). "Buying Supermajorities." *American Political Science Review* 90(2), 303–315.

Kreps, D. M. and R. Wilson (1982). "Sequential Equilibria." *Econometrica* 50(4), 863–894.

Kuran, T. (1991). "Now Out of Never: The Element of Surprise in the East European Revolution of 1989." *World Politics* 44(1), 7–48.

Kydland, F. E. and E. C. Prescott (1977). "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *Journal of Political Economy* 85(3), 473–491.

Ladha, K. K. (1992). "The Condorcet Jury Theorem, Free Speech, and Correlated Votes." *American Journal of Political Science* 36(3), 617–634.

Laffont, J.-J. and D. Martimort (1997). "Collusion under Asymmetric Information." *Econometrica* 65(4), 875–911.

Landes, W. M. and R. A. Posner (1975). "The Independent Judiciary in an Interest-Group Perspective." *Journal of Law and Economics* 18(3), 875–901.

Magaloni, B. (2006). *Voting for Autocracy: Hegemonic Party Survival and Its Demise in Mexico*. Cambridge University Press.

Moore, J. (1992). "Implementation, Contracts, and Renegotiation in Environments with Complete Information." In J.-J. Laffont (ed.), *Advances in Economic Theory: Sixth World Congress*, Vol. 1, 182–282. Cambridge University Press.

Nalepa, M. (2022). *After Authoritarianism: Transitional Justice and Democratic Stability*. Cambridge University Press.

North, D. C. and B. R. Weingast (1989). "Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in Seventeenth-Century England." *Journal of Economic History* 49(4), 803–832.

Ottaviani, M. and P. N. Sørensen (2001). "Information Aggregation in Debate: Who Should Speak First?" *Journal of Public Economics* 81(3), 393–421.

Persico, N. (2004). "Committee Design with Endogenous Information." *Review of Economic Studies* 71(1), 165–191.

Simpser, A. (2013). *Why Governments and Parties Manipulate Elections: Theory, Practice, and Implications*. Cambridge University Press.

Stephenson, M. C. (2003). "When the Devil Turns. . . : The Political Foundations of Independent Judicial Review." *Journal of Legal Studies* 32(1), 59–89.

Svolik, M. W. (2012). *The Politics of Authoritarian Rule*. Cambridge University Press.

Tirole, J. (1986). "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations." *Journal of Law, Economics, and Organization* 2(2), 181–214.