When Consensus Acquits

Capture and the defence of collective judgment

Torun Dewan

London School of Economics

Abstract

Collective judgment can be captured. A court, a jury, a committee can be swept or suborned to deliver the verdict a power wants. We distinguish two ways this is done. A bench can be forced to agree – the show trial, the acclamation, the manufactured unanimity. Or it can be packed to control the outcome while dissent goes on the record. The two threats call for different defences. Against the first, the optimal defence reads agreement against itself: a captured bench manufactures consensus, so consensus becomes the signature of capture, and the rule that defends the court refuses to convict on it. We characterise this defence and show its sharp form is a rule that looks perverse – a bench that convicts unanimously acquits the accused (Sanhedrin 17a). A family of devices follows as complements: a tilt toward acquittal, an order of voting that denies deference, benches that grow with the gravity of the charge. The Sanhedrin's capital procedure assembles them all, and we read it as the clearest instance of the design. Against the second threat, discrediting consensus does nothing; that bench needs entrenchment, not suspicion. Sorting the institutions of collective judgment by the capture they fear is the account's wider purpose.

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 the verdict the court discards; any surer route costs more than it is worth. 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. An optimal rule therefore 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 increases 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; there agreement turns from the strongest evidence into the weakest.

The mechanism is this. Suppose that with some probability the bench is compromised: it defers to a dominant member, or is directed to its verdict by a power that wants a conviction. A compromised bench then returns unanimous conviction whatever the truth. The compromise we model is *state-independent*: it concentrates on consensus regardless of guilt. This is the clean limit of, but not identical to, an informational cascade (Banerjee, 1992; Bikhchandani et al., 1992). A cascade's direction is set by early signals, so it stays correlated with the truth; it would leave even the unanimous count informative about guilt, and so would not deliver the clean inversion that is our first main result. 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 the cheap and reliable form of corruption manufactures the consensus the rule refuses to honour, while the routes that would instead manufacture a conviction – a fabricated dissent, a partial capture – are dearer than they are worth; so the rule deters the manipulation it cannot observe. Its force is a commitment; that 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: it withholds conviction from a unanimous bench but never imposes it.

One threat 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 – not as deductions from a single inequality but as the complements a single problem calls for: how to render judgment when the court may not be free. Each answers a particular route to capture, and each draws on a cost the procedure itself supplies, since recorded and reasoned dissent makes a fabricated dissent dear. The procedure manufactures the independence it then audits. It sizes itself to price capture out.

And it commits in advance to a rule it would be tempted to break, so that the manipulation it guards against is never tried. The unity is that one threat organises these devices, not that one primitive entails them all; each carries its own scope condition, stated where it enters.

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). These 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. A state-independent component concentrated on consensus drives the likelihood ratio at the top toward one, and can thereby invert the posterior in the count; this 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 at the top: what a conviction rule would read as the strongest case – stronger even than the lone dissent – becomes, under capture risk, weaker than that dissent and falls below the conviction threshold, so the optimal procedure acquits on it. The second is that the contamination is *one-sided*: it falls on the convict-consensus alone. The rule's suspicion is therefore asymmetric and tracks the direction in which courts are captured; 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; that commitment is the loss-minimising rule against a captor who best-responds. Around it the other features of the Sanhedrin – its pro-acquittal asymmetry, its graded benches, its order of voting – fall into place as the complements that threat calls for, each with its own institutional cost rather than a deduction from the rule itself. 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 belongs where a guaranteed verdict against a marked defendant must be deterred – the Sanhedrin, or the procedure one would build against the show trial. It is absent where capture works through a standing majority, as in a packed constitutional court.

2 Two threats to collective judgment

A body assembled to judge can be captured in two ways, and the two are not defended against alike.

A power can force the bench to agree. It sweeps the panel and seats its own, or suborns the judges it inherits, until the verdict it wants – a marked enemy condemned, or driven into exile

– is the verdict returned. The mark of this capture is unanimity. A directed bench does not divide; it concurs. The show trial ends in a single voice, the court convened to banish a rival returns its expulsion unopposed, the acclamation carries without a dissent, the politburo records none. The agreement is manufactured, and it is manufactured because agreement is what the power needs the verdict to display.

A power can instead pack the bench for a majority. It need not silence dissent; it need only outvote it, case after case, on a standing court whose business is a stream of decisions. The dissenters dissent, on the record, and lose. This is the capture of the constitutional courts of our own day – Poland after 2015, and the cases that followed it (Nalepa, 2022; Chiopris et al., 2025). Its mark is not unanimity but control, and control leaves the count looking ordinary.

The distinction is not a refinement; it sorts the defences. Against forced consensus a court can defend itself through the count, because the count betrays the capture: unanimity is the signature the directed bench cannot help leaving. A rule that refuses to convict on unanimity, a record that preserves dissent, an order of voting that denies deference – each reads agreement against itself, and each is useless against a power that never needed agreement. Against a directional majority the count says nothing; the bias is in who sits, not in how they vote. That bench is held off not by suspicion of its verdicts but by entrenchment – terms a packer cannot quickly fill, a jurisdiction he cannot quietly strip, a price on packing raised in advance. The insurance theory of judicial independence studies that defence: courts are built strong by power-holders who may later lose, as protection against the day they do (Ginsburg, 2003; Stephenson, 2003; Helmke, 2005; Vanberg, 2005). This paper studies the other defence.

Our model is of the first threat. We characterise the optimal response to a bench that may be forced to consensus, and we read the Sanhedrin's capital procedure as the institution that assembles it. The second threat we take as the boundary of the account, returning to it only to mark what our rule does not reach (§8). The Sanhedrin's rule is, in turn, one defence among several against the first threat. Once its parts are in hand we set the family side by side – immunity at the gate, composition at the bench, the verdict rule, the reversal – and order them by the capture each answers (§11).

3 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; Little, 2017); 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). We take that primitive in a sharp form: a captured bench returns a unanimous conviction whatever the facts. Looser forms of loyalty-staffing – majority control, a packed but not unanimous bench – fall outside this primitive, and we return to them when we mark the rule's scope.

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 (Condorcet, 1785) and its strategic heirs (Austen-Smith and Banks, 1996; Feddersen and Pesendorfer, 1998; McLennan, 1998) 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 (Ladha, 1992; Berg, 1993). 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.

4 Model

We cast the model as a court reaching a verdict. It applies to any body that must take a binary decision it cannot take on trust – a jury, a committee, a board – and that may be swept or suborned to the decision a power wants.

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 $k \in \{0, \ldots, n\}$.

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 independent 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. The procedure is built to confront it, not assumed to abolish it. The rate is a standing, exogenous background – a primitive of the polity, not an equilibrium object; in Section 8 we add a patron who chooses whether to manufacture a conviction *on top of* it, and show the rule deters that added manipulation, driving the patron's capture to zero while the background rate persists.

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; he 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 8. That regime 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.

5 Consensus is self-discrediting

Feddersen and Pesendorfer (1998) showed that requiring unanimity to convict can convict the innocent: under strategic voting a juror who must be pivotal reads guilt off his own decisiveness and votes to convict against his signal. Their unanimity is dangerous because honest jurors are pressed into it. Ours is dangerous because a captured bench manufactures it. The same event – a unanimous conviction – that their rule over-trusts, ours discards. Here is why.

For $k < n$ and $\lambda < 1$ 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 $\lambda = 1$ no free bench exists and every $k < n$ has probability zero under both states; the expression above is the value for all $\lambda < 1$ and the limit as $\lambda \uparrow 1$.) 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 $\{0,\dots,n-1\}$ the ratio L is strictly increasing, so $\{k \le n-1 : L(k) \ge \tau\} = [k_\tau, n-1]$ with $k_\tau = \min\{k : L(k) \ge \tau\}$, 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. □

The result conditions on $\tau \in \big(L(n), L(n-1)\big)$, and this window is a genuine restriction, not a consequence of capture and asymmetric costs alone. Capture with $\lambda > \lambda^\star$ secures $L(n) < L(n-1)$ – it opens the window – but where the cost ratio places τ within it is a separate matter. A threshold above $L(n-1)$ convicts nowhere; one below $L(n)$ convicts everywhere, unanimity included. The rule's distinctive verdict is the designer's when the stakes put τ between the two.

6 The Sanhedrin: the worked instance

The argument so far is general. A body must judge; its freedom is in doubt; the optimal rule reads its agreement against itself. We turn now to the institution that assembles the whole design. The Sanhedrin's capital procedure is the clearest instance of a court built against forced consensus, and its parts answer one for one to the model's.

Remark 1 (Which claims need which reading of the sugya). 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; 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 claims divide cleanly by how much of this they need.

Reading-robust. The non-monotone inference (Proposition 1), the acquittal at unanimity (Proposition 2), its optimality among count rules against a best-responding patron, the deterrence of forced consensus (Proposition 5), the directional asymmetry, and the robustness to staged dissent (Proposition 8) turn on the count's likelihood structure: independent informative behaviour in the free regime, which fixes $L(k)$ below unanimity, together with a state-independent capture atom on consensus, $k = n$. They do not ask *when* the count is taken, provided the opening tally carries that same structure – free-regime opening positions as informative as final votes, and the capture atom at unanimity. Granting this, whether k records the opening tally or the final one a captured bench places the same atom and the same non-monotone discount follows.

Process-dependent. The top-pivot Lemma 1, the every-equilibrium keystone (Proposition 3), the corollary reversing the pivotal incentive, the partial-capture results (Proposition 10), the bench-size threshold (Proposition 11), and the junior-first complementarity instead read the count as a bench's sequential, junior-first expression of position. These do *not* follow from the opening reading: Maimonides locates the trigger at the opening, where there is no observed prior tally for a last mover to be pivotal over, so the top-pivot mechanism would require a separate

model of opening statements or reason-giving that we do not supply. We claim them under the final-vote reading, and mark the dependence where each is incurred.

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 by forestalling a deference cascade – the independence mechanism, distinct from the standing capture rate λ, which the rule confronts rather than holds 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.

7 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 do not solve this sequential game in full – we leave the interior strategies, the informativeness of the prior votes, and uniqueness uncharacterised – but pin its outcome at the decisive node. 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 pools two sources – capture's manufacture and a free bench's own route to acquittal – and conviction is unwarranted on each: on the free path because the decisive judge there preferred acquittal, on the captured atom because $\lambda > \lambda^\star$ holds $L(n)$ below τ (Propositions 1–2). The step from the decisive judge's action to the count's verdict is this union over the two regimes, and the argument fixes it 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. (At the measure-zero posterior exactly at the threshold

both votes are optimal; we adopt the convention that the decisive judge dissents, which makes the keystone action unique and is vacuous generically, when the posterior is not exactly at τ.) 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 pins the action there to the sequential game. The senior's choice at the top pivot is forced in every equilibrium, on the path or off it, and that is all the keystone needs: wherever a free bench would complete a convicting unanimity, the decisive judge instead dissents. We make no claim that the top pivot is reached with positive probability – the keystone is a statement about the action at that node, not about how often it is visited.

Proposition 3 (The rule's keystone 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. A sequential equilibrium exists, and 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: the rule never converts a free bench's intended conviction into the acquittal at unanimity. We do not characterise the interior votes, claim that they are informative, or claim uniqueness; the deterrence result below rests on this top-pivot fact, and the optimality result on the posterior the count carries – this fact entering there only to show that a commitment cost vanishes.*

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 keystone needs nothing more: it holds in every sequential equilibrium of the game, whose existence was fixed at the outset. Whether the senior's vote in addition *tracks his signal* – responsiveness – is a separate matter, holding only when his two signal-contingent likelihood ratios straddle the threshold τ (itself a likelihood-ratio threshold, with the prior odds already folded in), $L_{\text{prefix}}\frac{p}{1-p} \geq \tau > L_{\text{prefix}}\frac{1-p}{p}$, with L_{prefix} the likelihood ratio he reads off the $n-1$ history; where they do not straddle it he is unresponsive at the pivot. Either way the top-pivot conclusion of Lemma 1 is unaffected – a free convicting bench still stops at $k = n-1$ – so we neither assert that a responsive equilibrium exists nor rely on one, and 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. At the top pivot a free bench's decisive judge who would convict dissents, holding the count at $n-1$, which the rule convicts (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 (Feddersen–Pesendorfer inverted)**.** *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). This is the exact inverse of Feddersen and Pesendorfer: the same pivotal logic, turned the opposite way – where their unanimity rule convicts the innocent, ours acquits. The machinery differs, theirs asymptotic and welfare-based, ours an inference at a finite top pivot, but the conclusion is the mirror image. The most celebrated case against unanimity and the case we make for it rest on one and the same pivot, read in opposite directions.*

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 no-reversal content – that the rule never converts a free bench's intended conviction into the acquittal at unanimity – is a *strategic*-voting result: under strategic voting a conviction-leaning free bench dissents to $k = n-1$ (Lemma 1) and is convicted, so its conviction is never reversed. Under *sincere* voting this does not hold – a genuinely unanimous free bench casts n convicting votes and is acquitted at $k = n$, a reversal of its intended conviction – and there it is the commitment-cost analysis of the deterrence section, not the no-reversal property, that carries the rule. Even under strategic voting we do not claim a free bench never reaches $k = n$: a decisive judge who prefers acquittal completes the unanimity, and there the rule acquits as he intends. Proposition 3 sharpens this from an outcome read off one vote into an equilibrium statement: the action at the decisive node is the one every sequential equilibrium forces, on the path or off it, so the lemma is no artefact of a convenient selection. We claim neither that the node is reached with positive probability nor that a responsive equilibrium exists; the keystone is the forced action there, which is all the deterrent needs. 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.

8 Capture and the asymmetry of the rule

So far the probability λ that the bench is compromised has been an exogenous background rate. We now add to it an endogenous actor, keeping the two distinct. The standing $\lambda > \lambda^\star$ remains – it is the structural risk that makes a unanimous count suspect at all – but on top of it a *patron* (a sovereign, a prosecutor, a powerful litigant) decides, case by case, whether to manufacture a conviction of his own. He 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 abstaining is normalised to a payoff of zero; he may *capture* the bench at cost $\kappa < B$. What the results below endogenise and drive out is this *added* capture, not the background λ: the rule leaves the patron no profitable manipulation, so the capture-induced wrongful-conviction rate falls to zero while the standing λ persists, neutralised. 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 votes back. A bare convicting majority can be peeled by that last move – a vote bought back drops it below the margin that convicts – so it guarantees nothing. This is the Groseclose and Snyder (1996) configuration: a coalition a counter-buyer can attack is secured only by overbuying toward a robust supermajority. The captor is thus pushed toward larger convicting coalitions, and a count he must defend against the counter-buyer carries a premium $\delta > 0$ over the blunt manufacture of agreement, larger the narrower the acquittal margin he must survive.

Proposition 5 (Capture deterrence)**.** *Fix the standing compromise rate $\lambda > \lambda^\star$ of Section 8 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.*

The patron cannot cheaply ride on it: the standing compromise is not his to direct, so to obtain a conviction he must establish his own control of the count – capture the bench (κ) and stage the dissent that leaves $k = n-1$ (δ) – a full route costing $\kappa+\delta > B$. The κ is his own cost of control, not a background cost sunk for him, so he cannot convert the standing consensus by paying δ alone, and cannot profitably turn a verdict out of the bench. 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 likelihood ratio at unanimity, $L(n)$, 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 $d : \{0, \ldots, n\} \to \{\text{convict}, \text{acquit}\}$ 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 $\delta > B - \kappa$.

Two features of the environment make a comparison over *all* count rules well-posed, rather than one confined to rules the captor happens to attack at unanimity. First, the patron's cost technology covers every count, not only $k = n$: blunt capture manufactures unanimity at κ, and by the Groseclose–Snyder remark above any count short of unanimity leaves a defensible margin and so costs the premium $\kappa + \delta$. The only count a patron can cheaply force is therefore $k = n$; no rule can be undercut at an interior count it relies on, so ranking count-measurable rules needs no separate cost for each count. Second, the standing compromise rate $\lambda > \lambda^\star$ is an exogenous background – the structural risk that makes a unanimous count suspect at all (Proposition 1) – and not the patron's choice; the patron is a separate, strategic actor whose *added* capture the rule deters. The optimality stated next is thus against a best-responding patron, holding that background fixed: the rule does not abolish λ but neutralises its effect at unanimity while pricing out the patron's addition. The two roles do not migrate – one is the environment the rule is designed for, the other the adversary it is designed against.

Proposition 6 (The rule is the optimal response to capture)**.** *We compare the count-measurable verdict rules a tallying court can commit to, a compromised bench producing only the all-convict profile. Among them:* (i) *every optimal rule convicts on the interior exactly where $L(k) \geq \tau$, that is on $[k_\tau, n-1]$, because by the cost technology above no interior count is profitably manufacturable, so counts $k < n$ are produced only by free benches and 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 separate the rules, by strategic equivalence rather than any pointwise comparison: under each rule's own equilibrium a free bench obtains the verdict its information warrants – a conviction-leaning bench convicts (at $k = n-1$ under the acquit rule by Lemma 1, at $k = n$ under the convict rule), an acquittal-leaning bench acquits – so it is correctly adjudicated under either rule and drops from the comparison. The captured atom alone decides it. 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 rules at $k = n$, and the reason is strategic equivalence, not a pointwise comparison holding behaviour fixed: we evaluate each rule under its own equilibrium. Under the acquit rule a free bench whose decisive judge would convict dissents to $k = n-1$ and is convicted (Lemma 1), so the free benches reaching $k = n$ are exactly those whose decisive judge prefers acquittal, whom the rule acquits – correctly. Under the convict rule that same conviction-leaning bench instead completes the unanimity and is convicted – also correctly, since absent capture the posterior at $k = n$ exceeds τ. Either way a free bench obtains the verdict its information warrants, at a different count under the two rules and at no error under either; we do not transplant the acquit-rule equilibrium onto the convict rule. Free benches thus drop from the comparison, and the captured atom alone separates the rules. Acquittal is optimal iff the wrongful convictions it removes on captured benches outweigh the guilty acquittals it incurs there, $(1-\pi)c_I\rho^\star > c_G\pi\rho^\star$, that is $\tau > 1$ – equivalently $c_I/c_G > \pi/(1-\pi)$, wrongful conviction costly enough relative to the prior odds of guilt, the maintained condition of Proposition 2 (the raw asymmetry $c_I > c_G$ suffices only when guilt is not a priori more likely than not); 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 concentrates 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: it falls below $L(n-1)$ at the inversion point $\lambda_c = \lambda^\star$ and, continuing to fall, crosses the decision threshold τ at a higher value $\lambda_c^\tau \geq \lambda^\star$ – higher because $\tau \leq L(n-1)$, so $L(n)$ reaches τ only after it has passed $L(n-1)$. Inversion is necessary but not sufficient: only for $\lambda_c > \lambda_c^\tau$ does $L(n) < \tau$ hold – the maintained $\tau \in (L(n), L(n-1))$ of Proposition 2, now read in λ_c – and there the optimal conviction set 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 $\lambda_c > \lambda_c^\tau$, 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 im-

plementation 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]

8.1 When capture concentrates 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 they mark the scope of the rule.

The first is the nature of the threat the rule is built against. The deterrence does not run through the voting margin, and cannot: a majority of two convicts, so on a bench of twenty-three a captor with thirteen convicting votes wins even if the other ten acquit, and a lone dissent against an otherwise unanimous conviction merely moves the count from n *into* the convicting band. What the rule answers is not a thin margin but the *instrument* of capture. A compromised bench does not deliberate; it concurs. The cheap and reliable way to guarantee the verdict of a marked defendant is therefore to manufacture that concurrence – unanimity – and it is this blunt capture (Section 8) that the rule discredits and refuses. To convict instead through a calibrated sub-unanimous count, the captor would have to fabricate dissent, which is dear, the cost δ; capture buys agreement, not a dial it can set to a bare majority. The rule's bite thus rests on capture being blunt – consensus-manufacturing – rather than finely tuned. Where a captor could cheaply seat a precise four-to-one, its special force reduces to the δ wedge, one fabricated dissent atop the generic cost of suborning a convicting majority. That is why the rule answers the guaranteed conviction of a marked defendant, where the threat is exactly manufactured consensus, and not the directional control of a docket, where a bare majority governs and dissent survives on the record.

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

[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.

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 any compromised distribution that places state-independent mass below unanimity, once $\bar{\varepsilon}$ is set count by count. At each convicted count $k \in [k_\tau, n-1]$ the contaminating atom pulls $L(k)$ toward 1, and conviction survives while that atom stays below the count's slack to τ, the slack converted to a mass by the factor $\lambda(\tau - 1)$; at unanimity the leak works the other way – it raises $L(n)$ – so acquittal survives only while the residual atom is large enough to keep $L(n)$ below τ. Taking $\bar{\varepsilon}$ as the tightest of these per-count bounds leaves the rule unchanged for any contamination within it; the single staged dissent above is the case in which only $k = n$ and $k = n-1$ bind. And the margin is exactly what the recorded-dissent machinery of this subsection secures: making a fabricated dissent dear is what holds ε small.

8.2 When capture is not blind to guilt

The inversion of Proposition 1 was derived for *state-independent* capture: the compromised bench returns unanimity at a rate that does not depend on whether the defendant is guilty. One might object that real capture is not perfectly blind – a power may find it cheaper, or a bench may defer more readily, when the evidence already leans toward conviction. The keystone tolerates this. It survives all capture that is blunt enough, and fails only when capture becomes a near-clean signal of guilt.

Let the bench be compromised at a rate that may differ across states: λ_G when the defendant is guilty, λ_I when innocent, with $\lambda_G \geq \lambda_I$ – capture, if anything, strikes the guilty more often. State-independence is the case $\lambda_G = \lambda_I$; the polar opposite, $\lambda_I = 0$, is a bench suborned only against the guilty. Write $\rho \equiv \lambda_I/\lambda_G \in [0, 1]$ for the bluntness of capture: $\rho = 1$ is fully blind, $\rho = 0$ a clean guilt signal. Below unanimity only free benches produce the profile, so

$$L(k) = \frac{1-\lambda_G}{1-\lambda_I}\left(\frac{p}{1-p}\right)^{2k-n}, \qquad k < n,$$

the ladder of Proposition 1 shifted down by the factor $(1-\lambda_G)/(1-\lambda_I) \leq 1$. At unanimity

both regimes contribute, now unequally:

$$L(n) = \frac{(1-\lambda_G)\,p^n + \lambda_G}{(1-\lambda_I)(1-p)^n + \lambda_I}.$$

Proposition 9 (Robustness to informative capture)**.** *Fix $\lambda_G > \lambda^\star$, p, and n. There is a threshold $\rho^\star \in [0,1)$ such that the top inversion $L(n) < L(n-1)$ holds for every $\rho \in (\rho^\star, 1]$ and fails for every $\rho < \rho^\star$. The acquit-at-unanimity rule of Proposition 2 is therefore the optimal verdict for all capture blunt enough, ρ above the (weakly higher) level at which $L(n)$ falls through τ. State-independence, $\rho = 1$, is sufficient but not necessary.*

Proof. As λ_I rises toward λ_G (that is, as $\rho \uparrow 1$), the denominator of $L(n)$ increases, so $L(n)$ falls; and $L(n-1) = \frac{1-\lambda_G}{1-\lambda_I}\,(p/(1-p))^{n-2}$ rises, since $1-\lambda_I$ shrinks. The gap $L(n-1) - L(n)$ is thus strictly increasing in ρ. At $\rho = 1$ it is positive (Proposition 1 with $\lambda = \lambda_G > \lambda^\star$). At $\rho = 0$ we have $L(n) = [(1-\lambda_G)p^n + \lambda_G]/(1-p)^n$ and $L(n-1) = (1-\lambda_G)(p/(1-p))^{n-2}$; since $p > \frac{1}{2}$ gives $p^2 > (1-p)^2$, the free part of $L(n)$ already exceeds $L(n-1)$, and the atom λ_G in the numerator only widens the gap – with no innocent ever railroaded, unanimity is the strongest evidence, not the weakest. By monotonicity and continuity the gap changes sign once, at a single $\rho^\star \in (0,1)$; the inversion holds above it. The same monotonicity carries $L(n)$ downward through the fixed threshold τ at a cutoff at least as high, above which the rule acquits at unanimity. □

The condition has a reading that matters for what the rule is *for*. Capture is blind to guilt exactly when the power wants a *particular* verdict – this defendant condemned, whether or not he did it. That is the directed bench of Section 2: forced consensus is state-independent almost by definition, so $\rho \approx 1$ and the inversion is robust precisely where the rule is meant to bite. Capture that strikes only the guilty – ρ small – is a different and milder thing, a bench that leans to conviction on strong evidence rather than one suborned to a foregone conclusion, and against it unanimity is informative and should be honoured. The rule does not defend against every departure from independence. It defends against the one a power seeking a particular conviction creates, and that one it tolerates in full.

8.3 Robustness to capture

Voting convention. This subsection is a robustness exercise conducted under *sincere* voting: a free judge votes his signal, or his posterior at his own information set, without strategising over how the downstream verdict rule will read the completed count. It is deliberately separate from the strategic-voting equilibrium analysed earlier. A free judge who reasoned strategically – anticipating that a completed unanimity acquits – might vote otherwise, and we do not solve that fuller game. What we claim is only the comparative-statics conclusion that no partial attack pays, and that conclusion turns on the captor's costs, not on the free judge's sophistication. Proposition 11, which prices the bench off these partial-capture success rates, inherits the same

convention.

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 10 (Partial capture fails)**.** *Suppose voting is junior-first (Sanhedrin 4:2), the partially compromised regime stands at rate $\mu > \mu^\star$, each judge is suborned at cost c, a credible fabricated dissent costs a further δ, and the bench is large enough that $pB \le (n-1)c$ and $(n-1)c + \delta \ge B$. Then no partial attack is profitable. A captor who leaves one judge free and suborns the other $n-1$ to convict relies on that judge to dissent and leave $k = n-1$. His success falls the later the free judge sits: a judge who sees more of the apparent consensus before him discounts it – exactly as the rule discounts unanimity – and, suspecting capture, completes the unanimity to trigger the acquittal. At the senior's seat the attack is self-defeating; the captor's best version leaves the* earliest *judge free, who has no prefix to discredit and dissents on an innocent signal, succeeding at rate at most p. It is priced out when $pB \le (n-1)c$. To convict with certainty the captor must instead* plant *a dissent: suborning the $n-1$ and seating one fabricated dissent caps the count below unanimity while the suborned convictions floor it in the band $[k_\tau, n-1]$, so the verdict is conviction whatever the free judge's seat or vote. This hybrid is position-independent and succeeds for sure, at cost $(n-1)c + \delta$; it is priced out, and full capture (dearer by one suborning) with it, when $(n-1)c + \delta \ge B$.*

Proof. Take the simple attack first: $n-1$ suborned to convict, one judge free, no fabricated dissent. A free judge who reaches a penultimate node sees $n-1$ convicting votes; as at $k = n$ in Proposition 1 the prefix is contaminated, a partially compromised bench – the only compromise he need weigh, conditioning on his own freedom – producing it with certainty. The prefix likelihood ratio

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

is strictly decreasing in μ by the affine-ratio argument of Proposition 1 and crosses τ at a unique $\mu^\star$ (the sequential analogue of $\lambda^\star$); for $\mu > \mu^\star$ it lies below τ, so a free judge at that seat, suspecting capture, completes the unanimity to trigger the acquittal except on his own erroneous signal, convicting at rate at most $1-p$. A free judge earlier in the order sees a shorter prefix and so a milder discount; the earliest judge sees no prefix to discredit. Under the sincere-voting convention of this subsection he votes his signal, and the route wrongfully convicts at rate p. That rate is a conservative bound, not an assumption the route needs: under strategic voting an early judge who weighs the standing rate μ knows that to dissent on an innocent signal is precisely to leave $k = n-1$ and convict the innocent, while to convict completes the unanimity and acquits – so a capture-suspecting judge dissents on an innocent signal less often, not more, and the wrongful-conviction rate is no higher than p. The simple attack's success is thus decreasing in the free judge's position, maximal at the earliest seat at rate p, with every intermediate seat between; the earliest is the captor's best and pricing it out, $pB \leq (n-1)c$, prices out all positions. To exceed rate p the captor must plant a dissent. Suborning $n-1$ and instructing $n-2$ to convict and one to dissent, he obtains $k = n-1$ if the free judge convicts and $k = n-2$ if he dissents, both in $[k_\tau, n-1]$ on a capital bench, so conviction follows whatever the free judge's seat or vote (and an adaptive instruction – convict if he dissents, fabricate otherwise – delivers $k = n-1$ exactly even when $k_\tau = n-1$); the attack is position-independent and certain. Its cost is the $n-1$ subornings and the staged dissent, $(n-1)c + \delta$, for a payoff $B - (n-1)c - \delta$, negative when $(n-1)c + \delta \geq B$. Full capture suborns one judge more for the same verdict, payoff $B - nc - \delta$, and is dominated. Abstention then dominates every route, 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. This is why the simple partial attack – suborn $n-1$ to convict, leave one judge free – weakens the later the free judge sits: a judge who sees the whole apparent consensus before him discounts it most and completes the unanimity to trigger the acquittal, so leaving the *senior* free is self-defeating, convicting only on his own error at rate at most $1-p$. The captor's best version leaves the *earliest* judge free, who has no prefix to discredit and votes his signal, dissenting at rate at most p; the bench size prices even this out, $(n-1)c \geq pB$.

A reader tallying assumptions should see the two wedges plainly: the gap δ between manufacturing consensus and manufacturing a *credible* dissent, and the per-judge cost c. The connection to the graded benches is already visible: the simple route is priced out by $(n-1)c \geq pB$ and the planted-dissent route by $(n-1)c + \delta \geq B$, both thresholds easier to clear the larger the bench.

Abler judges – a higher p – raise the simple route's prize pB and so strain that deterrent, which is why the bench must grow with the stake. The sequential, public order is essential: it is what lets a free judge see the apparent consensus before him and refuse to ratify it, collapsing the simple attack's success from the near-certainty a credulous bench would hand it.

The route that convicts for certain is to plant a dissent: suborn $n-1$ and instruct one to dissent, manufacturing a count in the convict band whatever the free judge does. 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. Full capture of all n is dearer by one suborning and dominated; the rule deters when the bench is large enough that $(n-1)c+\delta \geq B$ – blunt capture acquits, and disguised capture is too costly to disguise.

8.4 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 14, 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.

8.5 Capture for direction, capture for removal

Section 2 drew the two threats in the abstract. The contrast is sharpest in two real courts. Poland's Constitutional Tribunal after 2015 is the clearest contemporary capture: a governing party packed the bench – contesting appointments, seating loyalists, refusing to seat judges lawfully named before it – until it returned the rulings the government wanted (Nalepa, 2022; Chiopris et al., 2025; Sadurski, 2019). This is capture, real and consequential, but not the capture our rule answers. It is capture for *direction*: a reliable majority on a standing court, deciding a stream of cases, tolerating dissent on the record. A governing majority suffices, so nothing forces the bench to consensus, and the acquit-at-unanimity rule audits a signature this capture has no reason to produce.

The other branch is visible whenever a court is turned against a marked individual – to condemn him, or to drive him out. The directed verdict need not be a conviction; expulsion is the same instrument, a judgment that removes the man a power wants gone, and what the rule withholds it withholds from both. When a regime wants a *particular* adversary removed – an Old Bolshevik condemned in the Moscow trials, a purged officer before a military collegium, a senator exiled by a Roman treason court, a faction's enemy banished from a city-republic by a captured bench, as Dante was from Florence – the threat is exactly the one the Sanhedrin's procedure is built against: a guaranteed verdict against a marked defendant, secured the only cheap way a verdict can be guaranteed – by manufacturing the bench's concurrence. There capture does drive consensus. The tribunals returned their verdicts 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.

The purge pursues the same end without a courtroom. A ruler who commands both the apparatus that removes officials and the appointment of their replacements can purge them wholesale (Montagnes and Wolton, 2019); this is removal by administrative fiat, and no verdict rule reaches it. The show trial is that purge given judicial form – the removal routed through a court so that a verdict, and not merely a dismissal, certifies it. It is the court so used that our rule defends.

Ancient Athens shows the same design problem solved a different way. Ostracism was a purpose-built rule for expelling a political enemy: once a year the assembly could send a citizen

into ten years' exile, on a bare potsherd, with no charge and no defence.[3] It too was turned against rivals – Themistocles, Aristides, and Cimon all went out on the politics of the day – but its guards against capture were not the Sanhedrin's. Where the Sanhedrin discredits a manufactured consensus, Athens required breadth: a quorum of six thousand, so that no narrow faction could carry an expulsion alone, and a sanction kept mild and recoverable – ten years, no loss of property, no stain on the family – a light protection for a light penalty. The design held against the threat it feared and failed against the one it did not. When two of its leading men combined their followings to ostracise Hyperbolus, the broad quorum was met by a manufactured coalition; the rule was discredited and Athens never used it again. What the captured courts teach by their success, ostracism teaches by its collapse: an anti-capture device must be matched to the capture it faces.

The contrast fixes the theory's empirical content as a comparative, falsifiable prediction. Rules that refuse to convict on unanimity should appear in high-stakes, single-defendant, acquittal-protective procedures that record reasoned dissent – where the sanction removes a marked person, by death or by exile – and be absent from courts of review where a majority governs; the anti-capture apparatus of the latter takes the different form its different threat calls for – random assignment of panels, recusal, fixed terms, super-majority confirmation. A consensus-discrediting rule grafted onto a constitutional court would be a category error, and its absence there confirms the theory rather than contradicting it.

9 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 10). A simple partial capture – $n-1$ suborned to convict, one judge free – relies on the free judge to dissent, and the later he sits the more he discounts the apparent consensus and completes the unanimity to trigger the acquittal; the captor's best version leaves the earliest judge free and succeeds at rate at most p, at cost $(n-1)c$. To convict for certain he must *plant* a dissent – the *hybrid* – which secures a count in the convict band

[3] On the institution and its procedure see Aristotle, *Constitution of the Athenians* 22, with Forsdyke (2005); the ostracisms of Themistocles, Aristides, and Cimon are recounted across Plutarch's *Lives*, and the figure of six thousand at Plutarch, *Aristides* 7 (whether it was a quorum of all votes cast or of votes against the candidate is disputed). On the ostracism of Hyperbolus, secured by a combination of two leading men, after which the device fell from use, see Plutarch, *Nicias* 11 and *Alcibiades* 13, and Thucydides 8.73.

whatever the free judge's seat or vote, at cost $(n-1)c+\delta$. The captor takes whichever pays more, so the bench's worst case is $\max\{pB-(n-1)c,\ B-(n-1)c-\delta\}$. The hybrid dominates simple full capture, which buys one judge more for the same verdict, so the whole-bench route need not be costed separately.

Proposition 11 (The bench is sized to price out capture)**.** *A bench deters capture once*

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

a sufficient threshold increasing in the captor's stake B. And graver charges raise that stake: the value to a sovereign or faction of condemning this *defendant – silencing a rival, making an example, settling a succession – is far larger in a capital cause than in a money suit, so the private or political gain B from a secured conviction rises with the gravity of the matter, not merely the social cost of error. Graver matters therefore require larger benches: the grading $3 < 23 < 71$ is the bench scaling with the stakes of capture. The success rates it prices out are those of Proposition 10, computed under that subsection's sincere-voting convention.*

Proof. By Proposition 10 the simple partial attack – one judge free, the rest suborned to convict – succeeds at rate at most p (the earliest free judge is the captor's best choice, later seats only more suspicious), so it nets at most $pB-(n-1)c$, non-positive once $n \geq 1+pB/c$. The certain route is the hybrid: with one planted dissent the captor convicts with probability one at cost $(n-1)c+\delta$, nets at most $B-(n-1)c-\delta$, and is deterred once $n \geq 1+(B-\delta)/c$. Simple full capture, costing one more suborning, is dominated by the hybrid and so deterred with it. The captor takes whichever route pays more, so all are unprofitable once $n \geq 1+\max\{pB,\ B-\delta\}/c$; 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, and it increases in B. □

The hybrid in numbers. Take a per-judge cost $c=1$, a credible fabricated dissent at $\delta=2$, a signal accuracy $p=0.7$, and a captor who values conviction at $B=10$. Full capture buys all n judges and stages a dissent, at $nc+\delta = n+2$; it ceases to pay once $n+2 \geq B$, at $n \geq 8$. A designer who sized the bench against this route alone would stop at eight. But the hybrid leaves one judge free and buys only the other $n-1$, at $(n-1)c+\delta = n+1$ – one c cheaper, for the same certain conviction – and still pays at $n=8$, where $9<10$. It is deterred only at $n \geq 9$, where $n+1=10=B$. The simple partial attack, forgoing the planted dissent, succeeds at rate at most p and nets $pB-(n-1)c = 7-(n-1)$, dead by $n=8$, so it does not bind. The deterring bench is therefore the hybrid's, $n^* = \lceil 1+\max\{pB,\ B-\delta\}/c \rceil = 9$ – one judge more than full-capture accounting alone would suggest. Raise the stake to $B=22$ and the same arithmetic gives $n^*=21$: graver charges, larger benches.

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 under sincere voting, which vanishes as n grows provided p is bounded away from 1: on a bench of twenty-three a free-bench unanimity is then almost never seen, so the deterrent is almost costless to maintain. And by Lemma 1 a free bench that would convict stops at $k = n - 1$ rather than complete the unanimity, so the genuine loss is smaller still. 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, under standard regularity on G and a convex convening cost, an interior optimal bench that rises with the gravity of the cause – the grading as a smooth comparative static rather than a knife-edge.

10 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 the juniors' votes carry their own signals.

Proposition 12 (Junior-first and the unanimity rule are complements)**.** *In the sincere/deference benchmark of this section, junior-first voting yields independent signal votes, so an honest bench convicts unanimously with probability of order p^n, vanishing in n, whereas senior-first voting admits the deference cascade and 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. Strategic voting only sharpens the comparison, since by Lemma 1 a free convicting bench stops at $k = n - 1$ and*

reaches honest unanimity rarer still. 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 then convicts unanimously only if all n judges draw the guilty signal, 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 senior-first deference a captor need only corrupt the most senior judge, whom the rest copy. That alone sways the whole bench to convict – but a unanimous conviction is exactly what the acquit-at-unanimity rule throws out, so the shortcut buys the captor no verdict directly. Its bite is on the convicting route the rule does honour, $k = n - 1$: with the juniors copying a captured senior, the captor can assemble that near-unanimity by paying for one judge and staging a single dissent, so the cost of capture collapses toward a single c regardless of n and the grading of Proposition 11 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 10 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.

11 A family of defences

The Sanhedrin's rule is one defence among several against a single danger – a collective body turned against a marked enemy. They differ in where they act. A prosecution moves through stages: whether the case is brought at all, who sits to hear it, how the verdict is read, and whether it can be undone. A safeguard can be set at any of them.

Immunity acts first, at the threshold. It does not improve the verdict; it denies the court the case. Rome made its tribunes inviolable so that the magistrates could not haul them up at all,

and a legislator's immunity does the same today.[4] This is the earliest guard and the bluntest. It asks nothing of the bench and learns nothing from it: the protected man is neither convicted nor cleared, only left alone. Its strength is that it alone answers a capture whose weapon is the trial itself. When a power prosecutes to detain, to disgrace, to drain an opponent through a season of hearings, acquittal at the end repairs nothing – the punishment was the process. Only a case never brought escapes it. Dal Bó et al. (2006) show that shielding officials from prosecution can lower corruption and raise the calibre of those who enter public life – but only where the judiciary is weak enough to be turned against them, which is just when the gate-level guard is worth its cost.

Immunity defends only as a rule. The guard just described is fixed in advance and impersonal: it shields an office, not a favourite, and a power bent on capture cannot lift it for the case in hand. Protection granted after the fact is a different thing. A sovereign or a party leader may shield a particular ally once the charge has landed – declining to dismiss him, staying the inquiry, granting a pardon. This is not a defence against capture but an exercise of the same discretion that enables it, applied to protect an ally rather than to pursue an opponent: it favours the loyal, is withheld from others, and follows the protector's interest – the leader's calculus over a scandal-hit minister, which Dewan and Myatt (2007) model for the cabinet. Ex ante, the rule places the case beyond reach; ex post, discretion returns it to the authority the rule was meant to constrain.

Safeguards of composition act next, at the bench. Whether judges are appointed or drawn by lot, whether a faction can seat its own, decides if the bench can be packed before it ever votes.

The Sanhedrin's rule acts later, at the verdict. It lets the trial run and reads the result, withholding conviction from the consensus a captured bench would manufacture. This is the most discriminating of these defences. The evidence is heard, and a free bench may still convict the guilty by a margin short of unanimity; only the signature of capture is refused. Athens reached the same stage by a cruder route – a quorum so broad no narrow faction could carry the expulsion, and a penalty kept mild, rather than a vote read closely.

Recorded dissent acts last, after the verdict. A preserved minority opinion, and a ruling reversible only by a greater court, keep a captured verdict from being the final word.

The stages trade off. The earlier the guard, the more capture it forecloses and the blunter it is; the later, the more it discriminates, and the more it must let happen in order to discriminate. Immunity forecloses everything and learns nothing; the consensus rule forecloses little and learns most. Where the designer sets the defence turns on how far the court can be trusted to sit at all. Put this precisely. Let a guard be placed at $s \in [0,1]$, from the gate ($s = 0$, the case never brought) to the full trial decided at the verdict ($s = 1$). Running the process further yields information about guilt, $Q(s)$, increasing and concave with $Q(0) = 0$; it also leaves capture un-

[4] On the tribunes' *sacrosanctitas*, sworn with the creation of the office in 494 BC, see Livy 2.33, with Lintott (1999).

foreclosed, $R(s)$, increasing and convex with $R'(0) > 0$ – once a trial exists at all some capture is irreducible, the detention and disgrace of a process that is itself the punishment, and only foreclosure escapes it. Writing ν for the value the designer places on an accurate verdict and χ for the capturability of the court, the designer chooses s to maximise

$$W(s) = \nu\, Q(s) - \chi\, R(s).$$

Proposition 13 (Where to place the guard)**.** *The optimal placement satisfies $\nu Q'(s^\star) = \chi R'(s^\star)$ where it is interior. It is decreasing in the court's capturability χ and increasing in the value of accuracy ν. It is the gate, $s^\star = 0$ – immunity, which forecloses all and learns nothing – when $\chi/\nu \geq Q'(0)/R'(0)$, and the full trial decided at the verdict, $s^\star = 1$ – the consensus-discrediting rule, which learns most and forecloses least – when $\chi/\nu \leq Q'(1)/R'(1)$. The bluntness of the optimal defence, $1 - s^\star$, rises with χ/ν.*

Proof. $W'' = \nu Q'' - \chi R'' < 0$ since $Q'' < 0$ and $R'' > 0$, so W is strictly concave and the first-order condition $\nu Q'(s) = \chi R'(s)$ identifies the unique interior optimum where one exists. By the implicit function theorem $ds^\star/d\chi = R'(s^\star)/W'' < 0$ and $ds^\star/d\nu = -Q'(s^\star)/W'' > 0$. The corner $s^\star = 0$ obtains when $W'(0) = \nu Q'(0) - \chi R'(0) \leq 0$, that is $\chi/\nu \geq Q'(0)/R'(0)$; the corner $s^\star = 1$ when $W'(1) \geq 0$, that is $\chi/\nu \leq Q'(1)/R'(1)$. As $s^\star$ falls in χ/ν, $1 - s^\star$ rises in it. □

The family is thus ordered by a single ratio. Where the court can be trusted and accuracy is prized, χ/ν low, the designer runs the full trial and guards only the verdict – the surgical defence this paper has characterised. As the court grows more capturable the optimal guard moves earlier and blunter, through the bench and toward the gate, until at high χ/ν the case is best never brought at all. Each defence is the answer to a court trusted a little less; the rule we study is the one for a court that can still be trusted to sit, and the same logic orders the rest – match the guard to the capture it faces.

12 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. That correlation can break the count's monotonicity is not new – it is the lesson of his correlated votes. What is new is the particular form the correlation takes: it is *one-sided* and *state-independent*, concentrated on the convict-consensus a captor manufactures whether or not the defendant is guilty, and this is what inverts the posterior at the very top, where the symmetric correlation Ladha studies does not. New too is 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: our deterrence results read the acquit-at-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-to-convict* rule by letting jurors communicate before voting – so that, when their thresholds are sufficiently aligned, sincere voting is an equilibrium and that rule minimises error – we defend the opposite

rule, which 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.

The deterrent belongs, finally, with the political economy of judicial independence. That literature explains why a power-holder would build a court able to rule against him: an independent judiciary is insurance, valued by those who may lose office against the day they do, and sustained as the device through which rival powers enforce mutual restraint over time (Ginsburg, 2003; Finkel, 2008; Landes and Posner, 1975; Stephenson, 2003); courts under pressure defect when that support is absent (Helmke, 2005) and hold when a public will defend them (Vanberg, 2005), and how far that guarantee is realised varies markedly across the courts of a region (Helmke and Ríos-Figueroa, 2011). That work asks why a court is *empowered*. We ask the next question: granted a court whose verdicts can bind, how does its procedure keep those verdicts from being dictated by the very power that empowered it? The insurance theory secures the court's independence from without – terms, appointments, public support; our rule defends the

verdict from within, by refusing to honour the consensus a captor would manufacture. The two are complementary defences of the same good, and the two threats of Section 2 mark the division of labour: entrenchment answers capture by majority, a consensus-discrediting rule answers capture by forced agreement.

13 Conclusion

A court cannot see whether its judges are free. It observes the votes, not the hands behind them. We have asked what a court should do when it cannot tell a free verdict from a forced one, and the answer is to read agreement against itself. Where consensus can be manufactured, consensus is the signature of capture, and the rule that defends the court withholds conviction from it. The Sanhedrin's capital procedure is the clearest instance: a bench that convicts unanimously acquits, and the rest of the procedure – the tilt toward acquittal, junior-first voting, benches that grow with the charge – follows as complements, each earning its keep against a particular route to the directed verdict.

The account sorts the institutions of collective judgment by the capture they fear. A body exposed to forced consensus should adopt devices that discredit agreement; a body exposed to capture by majority should not, for such devices would buy it nothing. This is a claim one can take to other institutions. Where a court records and reasons its dissents, it makes a fabricated dissent costly and a manufactured unanimity rare – the constitutional courts that publish minority opinions hold, in our terms, a standing defence against forced consensus. Where a collegial body fixes an order of speaking that denies deference to its seniors, it protects the independence of the junior voice. Where a rule demands a supermajority, the question our account poses is whether it protects a blocking minority or empowers a controlling one: the first guards against capture, the second invites it. And where capture runs instead through packing, we expect to find not suspicion of the count but entrenchment – the defences the insurance theory of judicial independence describes. The two families of device should appear where the two threats do.

We have read the Sanhedrin as a design, though its capital court tried no one within the reach of the law that records it: the procedure is doctrine, carried by the Mishnah and by Maimonides, and not a docket. The claim here is therefore analytic, not historical. The procedure encodes the defence a body builds when it may be captured, and the encoding is legible whether or not the body ever sat. That a tradition reasoned its way to this design, and preserved the reasoning, is the evidence we read. The logic it records binds wherever judgment is collective and capture is possible.

A A repeated-game foundation for the deterrent

Under sincere voting the rule is in the main ex-post optimal (Section 8), 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 14 (The deterrent is self-enforcing)**.** *Evaluate the court's incentive at the count it observes, under the belief that makes acquittal hardest to sustain: a sitting panel that, unable to read the standing capture rate off its single case, takes its own unanimity to be genuine and would convict on it. (With correct beliefs the posterior at unanimity is $L(n) < \tau$, acquittal is statically optimal, and no temptation arises.) For that worst-case panel, 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*

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

g being that worst-case one-period temptation; for any weaker belief the inequality is slack, so $\beta^\star$ is a conservative sufficient threshold. The tempting belief itself arises 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, with the verdict evaluated at the court's information – the count – not at any hidden classification of the bench. Given correct beliefs the prescribed verdict is statically optimal at every count: on the interior the increasing-threshold verdict, and at $k = n$ the acquittal, since the posterior there is $L(n) < \tau$ under the maintained $\lambda > \lambda^\star$ and $\tau \in (L(n), L(n-1))$. With correct beliefs, then, no history tempts and the profile is self-enforcing for any β. Commitment binds only for the panel that believes its own unanimity genuine: there acquitting forgoes the one-period gain g – the worst-case temptation, a panel certain of its unanimity – while convicting is observed and triggers reversion forfeiting the deterrent D in every future period, so acquittal is sustained iff $g \leq \frac{\beta}{1-\beta} D$, that is $\beta \geq g/(g + D)$. For any weaker belief the temptation is smaller and the inequality slack, so $\beta^\star$ is a conservative sufficient threshold. The verdict record is public, so reversion fires on observed play and the off-path beliefs supporting it are trivial. 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.

Berg, S. (1993). "Condorcet's Jury Theorem, Dependency Among Jurors." *Social Choice and Welfare* 10(1), 87–95.

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.

Dal Bó, E., P. Dal Bó, and R. Di Tella (2006). "'Plata o Plomo?': Bribe and Punishment in a Theory of Political Influence." *American Political Science Review* 100(1), 41–53.

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

Dewan, T. and D. P. Myatt (2007). "Scandal, Protection, and Recovery in the Cabinet." *American Political Science Review* 101(1), 63–77.

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.

Finkel, J. S. (2008). *Judicial Reform as Political Insurance: Argentina, Peru, and Mexico in the 1990s*. University of Notre Dame Press.

Forsdyke, S. (2005). *Exile, Ostracism, and Democracy: The Politics of Expulsion in Ancient Greece*. Princeton University Press.

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.

Ginsburg, T. (2003). *Judicial Review in New Democracies: Constitutional Courts in Asian Cases*. Cambridge University Press.

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.

Helmke, G. (2005). *Courts under Constraints: Judges, Generals, and Presidents in Argentina*. Cambridge University Press.

Helmke, G. and J. Ríos-Figueroa (eds.) (2011). *Courts in Latin America*. Cambridge University Press.

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.

Lintott, A. (1999). *The Constitution of the Roman Republic*. Oxford University Press.

Little, A. T. (2017). "Propaganda and Credulity." *Games and Economic Behavior* 102, 224–232.

McLennan, A. (1998). "Consequences of the Condorcet Jury Theorem for Beneficial Information Aggregation by Rational Agents." *American Political Science Review* 92(2), 413–418.

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.

Montagnes, B. P. and S. Wolton (2019). "Mass Purges: Top-Down Accountability in Autocracy." *American Political Science Review* 113(4), 1045–1059.

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.

Sadurski, W. (2019). *Poland's Constitutional Breakdown*. Oxford University Press.

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.

Vanberg, G. (2005). *The Politics of Constitutional Review in Germany*. Cambridge University Press.