When Consensus Acquits

Capture and the design of the Sanhedrin's capital procedure

Torun Dewan

Preliminary working note, June 19, 2026. Comments welcome.

Abstract

Courts can be captured. A sovereign, a prosecutor, or a powerful litigant who wants a conviction can sweep or suborn the bench, and a procedure for rendering judgment must reckon with the possibility that its own judges are not free. We model the Sanhedrin's capital procedure as a design for exactly this problem. Its centrepiece is a rule that looks perverse – a court that convicts unanimously acquits the accused (Sanhedrin 17a) – and we show it is the optimal response to capture. Because a captured bench manufactures consensus, unanimity is the signature of capture, and a rule that refuses to convict on it deters the manipulation it cannot observe: the only verdict capture can force is the one the court will not honour. The rest of the procedure follows from the same primitive. The pro-acquittal asymmetry mirrors the direction in which courts are captured; the junior judges vote first, to manufacture the independence the rule audits and to deny a captor the shortcut of suborning the senior alone; and the bench grows with the gravity of the charge because a larger court prices capture out. The argument rests on commitment – the rule is time-inconsistent, and we locate its credibility in the rigidity of codified procedure, the bar on revision except by a greater court, and the reputational logic of repeated cases. The device is the one Solomon uses and the implementation literature studies: an off-equilibrium threat that makes manipulation unprofitable, so that on the path it need never be used.

1 Introduction

A collective decision to convict rests on the judgments of many, and the law must turn their votes into a verdict. The standard treatment of this problem assumes the judgments are conditionally independent. The count of convicting votes is then a sufficient statistic for guilt, the posterior is increasing in it, and the optimal rule is a threshold: convict when the count is high enough (Condorcet, 1785; Austen-Smith and Banks, 1996; Feddersen and Pesendorfer, 1998). More agreement is always more reason to convict.

An old rule cuts against this. The Mishnah grades the bench by the gravity of the charge and tilts every step of capital procedure toward acquittal (Sanhedrin 4:1–4:2): a bare majority acquits but a majority of two is needed to convict; the court opens for acquittal; a judge may change his vote to acquit but not to convict. The same mishna sets these rules opposite monetary

procedure, which decides either way by a bare majority of one; the asymmetries are reserved for capital cases alone, where the cost of the two errors diverges. The asymmetry is not the Mishnah's invention but its reading of Scripture – *lo tihyeh aharei rabim le-ra'ot*, "do not follow a multitude to do evil" (Exodus 23:2), taken to mean that for harm, a conviction, a bare majority will not serve (Sanhedrin 2a) – so the differential cost of the two errors enters the law as a verse, not a preference. The Talmud then states a rule that no monotone threshold can express. Rav Kahana teaches that a bench convicting to a man frees 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). The most agreement of all produces no conviction at all.

We show the rule is what a court chooses when it may be *captured*. Suppose that with some probability the bench is compromised – swept by a cascade (Banerjee, 1992; Bikhchandani et al., 1992), deferring to a dominant member, or directed to its verdict by a power that wants a conviction – and that a compromised bench returns unanimous conviction whatever the truth. The 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 possibility that the vote carried no information at all. A single dissent, by contrast, is almost impossible under capture, so it certifies that the court was deliberating and that all but one of its independent judgments pointed to guilt. The lone dissenter does not weaken the case. He authenticates it. Remove him and one can no longer tell a convinced court from a captured one.

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

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

Reading a procedure of the Talmud as a designed response to a strategic problem places this paper within a small program of such readings, after Aumann and Maschler (1985).[1] The contribution to the theory of collective judgment is a single primitive – a latent probability that independence has failed – under which the optimal verdict is non-monotone in the vote, and consensus is treated as suspect rather than decisive.

[1]The program reads sugyot not as solutions to problems the analyst poses, but as institutions: procedures, precisely legislated and defended in argument, that respond to strategic problems of information, commitment, and division. A companion paper develops the program at length.

2 Model

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

With probability $1 - \lambda$ the bench is *deliberative*: judges receive conditionally independent signals and vote 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 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.

3 Consensus is self-discrediting

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

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

At unanimity both regimes contribute:

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

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

Proof. For $k < n$ the compromised regime contributes nothing, so $\Pr(k \mid \omega) = (1-\lambda)\binom{n}{k}\rho_\omega^k(1-\rho_\omega)^{n-k}$ with $\rho_G = p$ and $\rho_I = 1-p$; the binomial coefficients cancel, giving $L(k) = (p/(1-p))^{2k-n}$, strictly increasing because $p > \frac{1}{2}$, and in particular $L(n-1) = (p/(1-p))^{n-2}$. At $k = n$ both regimes contribute: $\Pr(n \mid G) = (1-\lambda)p^n + \lambda$ and $\Pr(n \mid I) = (1-\lambda)(1-p)^n + \lambda$ are affine in λ, so $L(n)$ is a ratio of affine functions and its derivative has the constant sign of $ad - bc$ with $a = 1-p^n$, $c = p^n$, $b = 1-(1-p)^n$, $d = (1-p)^n$; here $ad-bc = (1-p)^n - p^n < 0$, so $L(n)$ is strictly decreasing on $[0,1]$, from $(p/(1-p))^n$ at $\lambda = 0$ to 1 at $\lambda = 1$. As $1 < L(n-1) < (p/(1-p))^n$ for $n \geq 3$ – and a capital bench has $n \geq 23$, so the bound binds with room to spare – the equation $L(n) = L(n-1)$ has a unique root $\lambda^\star \in (0,1)$, with $L(n) < L(n-1)$ for $\lambda > \lambda^\star$. Since L is increasing on $\{0, \dots, n-1\}$, the likelihood ratio then peaks at $k = n-1$. □

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

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

4 Reading

The rule conditions the verdict on the vote and on its credibility. A near-unanimous bench with one dissent is the most incriminating outcome the procedure admits, because dissent is the mark of a free court. Unanimity is exactly what a captured court produces, so the law refuses to convict on it. This is the sense the codifiers give the acquittal. Maimonides states it as settled law: a Sanhedrin all of whom open the capital case by convicting – *she-patchu kullam. . . techilah ve-amru kullan chayav* – acquits, for a capital bench must hold among it those who would argue the defendant's merit, and here there were none (*Mishneh Torah*, Hilchot Sanhedrin 9:1). The rule passes from the Bavli (Rav Kahana, Sanhedrin 17a) into the code but travels no further: the *Shulchan Arukh* does not carry capital procedure at all, the courts that administered it having lapsed with ordination, so the forward chain ends at Maimonides rather than at the later codes – the law is preserved as doctrine, not as a live docket. In the model the deliberative regime would almost surely have produced a defender; the absence of one 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 deliberative regime informative and holds λ down. The graded benches and the majority of two to convict set the asymmetric price of error that the threshold τ records. The rule that consensus acquits is the audit; junior-first is the safeguard that makes the audit meaningful.

5 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.

The rule introduces a decisive event at the top of the count: a judge whose vote moves the bench between $k = n - 1$ and $k = n$. There, voting to convict makes the bench unanimous and acquits, while dissenting holds the count at $n - 1$ and convicts. The pivotal judge's incentive is thus inverted at the top – he dissents to convict, and concurs to acquit.

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

Lemma 1 (Pivot structure)**.** *In the sequential game a judge is pivotal at exactly two counts: a* bottom *pivot, where his convicting vote carries the count into the conviction set, and the* top *pivot $c = n-1$, where his convicting vote completes unanimity and acquits. A judge who observes the running count knows which he faces, and at the top he supports conviction by dissenting. His posterior on guilt there, conditional on the $n - 1$ recorded convictions and his own signal, inherits the λ-discount of $L(n)$ in Proposition 1; so for $\lambda > \lambda^\star$ an innocent signal leaves him short of the conviction threshold and he concurs, completing unanimity and acquitting, while a guilty signal makes him dissent and convict.*

Proof. The verdict convicts iff $k \in [k_\tau, n-1]$, so a vote matters only where the two final counts it allows straddle a boundary: at $c = k_\tau - 1$ (convict $\to$ conviction, acquit $\to$ acquittal) and at $c = n - 1$ (convict $\to k = n$, acquittal; dissent $\to k = n - 1$, conviction). At interior counts both continuations lie in the conviction set, so no judge is pivotal. With observed votes the judge knows c. At the top his preferred verdict is conviction iff his posterior odds clear the cost-implied threshold; conditional on the $n - 1$ recorded convictions and his signal those odds are the discounted $L(n)$ of Proposition 1, which falls below $L(n-1)$ once $\lambda > \lambda^\star$, so an innocent signal puts him below the threshold and a guilty signal above it. □

Proposition 3 (Strategic voting: the rule binds only on captured benches)**.** *Under the rule of Proposition 2 and strategic voting, the judge who casts the completing vote on a bench standing at $k = n - 1$ obtains his preferred verdict: he dissents (yielding $k = n - 1$, conviction) when his posterior favours conviction, and concurs (yielding $k = n$, acquittal) when it favours acquittal. Consequently the rule overturns no deliberative verdict – a deliberative bench that would convict returns $k = n-1$, while a deliberative bench returning $k = n$ is one whose decisive judge favours acquittal – whereas a captured bench, forced to $k = n$ irrespective of the judges' information, is acquitted. Acquittal at unanimity binds only on captured benches.*

Proof. At the completing vote the count stands at $n - 1$ and the judge is decisive: by Proposition 2 a convict vote yields $k = n$ and acquittal, a dissent yields $k = n - 1$ and conviction. A payoff-maximising judge selects the vote whose verdict he prefers – dissent if his posterior

favours conviction, concurrence if it favours acquittal – so the realised verdict equals his preferred outcome. Hence conviction (at $k = n - 1$) occurs exactly when the decisive judge favours it, and $k = n$ occurs only when he favours acquittal, where acquittal is also his preferred verdict; in neither case is a deliberative bench's judgment reversed. A captured bench returns $k = n$ by construction, independent of any judge's information, and is acquitted. □

Corollary 1 (Inversion of the Feddersen–Pesendorfer incentive)**.** *Under the rule that requires unanimity to convict (Feddersen and Pesendorfer, 1998), the decisive event – casting the last convicting vote – is itself evidence of guilt, since it arises only when all others convict; a judge with an innocent signal may then rationally convict, and strategic voting fails toward wrongful conviction. Under the rule that acquits on unanimity the same pivotal reasoning runs the other way: completing unanimity triggers acquittal, so no judge completes it to convict, and strategic voting purifies the consensus signal rather than polluting it. The two rules invert each other's incentives because one acts on unanimity and the other withholds from it.*

Proof. Under the require-unanimity rule a juror is decisive only when all others vote to convict; conditioning on that event raises his posterior on guilt, so a juror with an innocent signal may rationally convict, and informative voting degrades toward conviction (Feddersen and Pesendorfer, 1998). Under the acquit-at-unanimity rule the decisive completing vote instead triggers acquittal, so by Proposition 3 no judge casts it to convict; the incentive is reversed. □

The dilemma of who casts the dissent now dissolves. Obtaining exactly $n - 1$ convictions looks like a coordination among the guilty-believing judges, each preferring that the conviction stand but that another cast the dissent. Sequence removes the coordination. Only the final vote can carry the count from $n - 1$ to n, so the last mover is the residual claimant at the top pivot: by Lemma 1 he alone decides whether a near-unanimous bench convicts or acquits, and he dissents precisely when his signal favours conviction. The judges before him are never at the top pivot – on a deliberative bench the event that all others convict has probability of order p^n – so the acquit-at-unanimity rule does not distort their votes and they vote their signals. Under junior-first (Sanhedrin 4:2) the residual claimant is the most senior judge: the order that protects independence also designates, with no need for a volunteer, who must break the self-defeating consensus. The two rules interlock: junior-first manufactures the independence that unanimity-acquits audits.

Remark 1 (The order is not innocuous). For a monotone rule the informative equilibria of sequential and simultaneous voting coincide, and the order of voting is irrelevant to what the bench learns (Dekel and Piccione, 2000). The acquit-at-unanimity rule is non-monotone – it withholds conviction at the very top of the count – so that irrelevance fails and 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 2 (Which equilibrium). The game also has uninformative equilibria – if no judge's vote is ever pivotal the bench may acquit regardless – which we set aside by attending to responsive play, as is standard (Feddersen and Pesendorfer, 1998). Among responsive equilibria interior judges are non-pivotal and so indifferent, while at each pivot a judge's posterior is monotone in his own signal, making informative voting his best response there; crucially the dissenter's identity is fixed by the order and not by the equilibrium, so it is robust across the responsive equilibria. We do not claim uniqueness, which the jury literature does not deliver.

6 Capture and the asymmetry of the rule

So far the probability λ that the bench is compromised has been exogenous. We now let it be chosen. A *patron* – a sovereign, a prosecutor, a powerful litigant – gains $B > 0$ from the conviction of this defendant whatever the truth, and may capture the bench at cost $\kappa < B$. Capture is *blunt*: it manufactures agreement, returning unanimous conviction $k = n$. To manufacture instead a particular interior count – in particular the lone dissent that would leave $k = n - 1$ – costs a further δ, because a captured bench must orchestrate a holdout it has no reason to seat. We take δ large: producing consensus is cheap, producing credible disagreement dear. This is the central assumption. 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.

Proposition 4 (Capture deterrence)**.** *Under any rule that convicts at $k = n$, the patron captures whenever $B > \kappa$, and the bench is compromised with positive probability. Under the rule of Proposition 2, which acquits at $k = n$, capture delivers acquittal, and faking a dissent to reach $k = n - 1$ costs $\kappa + \delta > B$; the patron therefore strictly prefers not to capture. In equilibrium $\lambda = 0$ and capture is deterred: the rule is manipulation-proof.*

Proof. If the rule convicts at $k = n$, capturing buys conviction (gross B) at cost κ, profitable whenever $B > \kappa$, so the patron captures and $\lambda > 0$. Under the rule of Proposition 2 the patron's only options are blunt capture, which yields $k = n$ and acquittal at net $-\kappa$, or capture with a fabricated dissent, which yields $k = n - 1$ and conviction at net $B - \kappa - \delta$. With δ large, $B < \kappa + \delta$, so both are loss-making and abstaining (payoff 0) dominates; hence $\lambda = 0$. □

The deterrence is sustained off the equilibrium path. Under sincere voting, because capture does not occur, a unanimous verdict would in fact be deliberative and convicting on it would be optimal ex post; it is precisely the committed refusal to convict that keeps capture away. The rule is then time-inconsistent, and its value lies entirely in the threat. This is a reason for procedure to be *rigid*: a rule fixed in advance and beyond revision in the individual case

can commit where discretion cannot, and rigidity is here the source of the deterrent rather than an administrative convenience. Under sincere voting the cost of the commitment is the occasional acquittal of a genuinely unanimous guilty bench, of order πp^n, a price worth paying when wrongful conviction is the graver error. Under strategic voting even this cost lapses: by Proposition 3 a deliberative bench that would convict returns $k = n-1$, and a unanimous count signals a bench content to acquit, so acquittal is optimal ex post and the deterrent needs no commitment at all. Commitment is the safeguard for the conservative, sincere-voting case, and the procedure supplies it.

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 5 (The asymmetry mirrors the threat)**.** *If $\lambda_c > \lambda_a$, unanimity is discounted only on the conviction side: the optimal verdict acquits at $k = n$ but does not convict at $k = 0$. 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 and the costs to reverse, the rule would invert. The direction of the rule's suspicion is the direction of capture.*

Proof. With two-sided capture the likelihood ratio at $k = n$ carries a contamination term increasing in λ_c, and at $k = 0$ a term increasing in λ_a (each capture type loads on its own extreme). When $\lambda_c > \lambda_a$ the discount to the ratio is larger at $k = n$ than at $k = 0$, so by the argument of Proposition 1 the optimal conviction set excludes $k = n$; at the acquittal end no symmetric force makes $k = 0$ trigger conviction, and the asymmetric costs disfavour conviction further, so $k = 0$ is not convicted. The conviction-side discount grows in λ_c/λ_a and in the cost of wrongful conviction; reversing both reverses the inequality and the rule. □

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

6.1 Robustness to capture

A captor seeking the count $k = n - 1$ that the rule still honours has two routes. He may leave one judge uncaptured and hope the verdict falls there, or capture the whole bench and fabricate a dissent. The first route turns on what the free judge infers, because the captor chooses which judge to leave free and the voting order is fixed.

By Lemma 1 a free judge who sees the count steers the verdict to his signal; the danger is a free judge who cannot. An early-moving junior votes before any conviction is recorded, so, voting blind, he dissents on an innocent signal, leaves the count at $n-1$, and convicts. The captor would therefore like to leave an early judge free. What stops him is that doing so is dear.

Proposition 6 (Partial capture fails)**.** *Suppose the captor corrupts $n-1$ judges and leaves one free, under junior-first voting (Sanhedrin 4:2), and let the cost of corrupting a judge rise in his seniority. Leaving the senior free is the cheap attack: by Lemma 1 the senior, moving last and seeing the count, steers the verdict to his own signal, so it convicts an innocent only when the senior errs, at rate $1-p$, for the price of the $n-1$ juniors. Leaving an early junior free raises the success rate to p – the junior, voting blind, dissents on an innocent signal – but requires corrupting the senior ranks in his place. If the resulting seniority premium in corruption cost exceeds $(2p-1)B$, the captor prefers the cheap attack, and partial capture convicts an innocent only at rate $1-p$.*

Proof. By Lemma 1 a free judge who observes $c = n-1$ secures his preferred verdict, so leaving the senior free yields conviction only on his own error, probability $1-p$. A free junior moving first does not observe the count; voting his signal he dissents on an innocent signal and leaves $k = n-1$, a conviction, so leaving a junior free yields conviction with probability p against an innocent defendant. Each attack corrupts $n-1$ judges, but the second substitutes senior corruptions for junior ones; let Δ be the resulting cost premium. The captor's gain from the dearer attack is the higher success rate, worth $\big(p-(1-p)\big)B = (2p-1)B$. He therefore prefers to leave the senior free whenever $\Delta > (2p-1)B$, and the operative partial attack convicts at rate $1-p$. □

The cost gradient carries the result. It is worth recording that the rule's own logic seems to point the same way, though less securely.

Remark 3 (An inference-based conjecture). One is tempted to dispense with the gradient. A free junior who reasoned about capture might discount an apparent unanimity as the senior does: conditioning on all others convicting, he should weigh that this is the configuration capture manufactures, and invert. The argument is not secure. He casts a single vote that must serve at both the bottom and the top pivot, and cannot tailor it to the one he turns out to face; and the weight he puts on capture in the suspicious event is an off-path belief, held only if the patron's propensity is fixed and bounded away from zero – a belief that a consistency refinement, reading an all-convict prefix as independent guilty votes rather than coordinated capture, may drive to zero. We record the inference channel as a conjecture and rest the result on the cost gradient.

So the operative partial attack convicts at rate $1-p$ and corrupts the $n-1$ juniors – the per-judge cost c of the next section is theirs – so partial capture is ineffective and dear precisely when judges are able (p high) and the bench is large, a first connection between the rule and

the graded benches. The sequential, public order is essential: it is what lets the senior, moving last, see the count and steer.

The remaining route is to capture all n and instruct one judge to dissent, manufacturing $k = n - 1$ directly. This is the only place the cost δ of a fabricated dissent is required, and the procedure makes it positive by design. The Talmud records dissent with its reasons (*Eduyot* 1:5), so a fabricated dissent is not a silent vote but a reasoned minority opinion that must survive examination, and the planted dissenter is a co-conspirator who may defect. The rule deters when $B < \kappa + \delta$: blunt capture acquits, and disguised capture is too costly to disguise.

6.2 Commitment and the foundations of the deterrent

Under sincere voting the deterrent asks the court to acquit a unanimous bench even when the unanimity is genuine, which is optimal ex ante but not ex post: once capture is deterred and $\lambda = 0$, the court would prefer to convict. The threat is then credible only if the court can bind itself, 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. 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. The value of a standing deterrent exceeds the one-time gain from convicting a single genuinely-unanimous defendant, so the rule is self-enforcing in the manner of the folk theorems: the shadow of future cases, not a one-shot promise, is what holds the court to its rule.

7 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 attack, partial capture (Proposition 6): its operative form leaves the dearer senior free and corrupts the $n-1$ junior judges, at per-junior cost c, so it costs $(n-1)c$ and convicts the innocent with probability $1-p$, when the free senior errs; the captor attacks only if $(1-p)B > (n-1)c$, where B is his gain from a conviction. The seniority gradient of Proposition 6 enters only in selecting this attack – it is what makes leaving the senior free cheaper than leaving a junior free – so c here is the junior cost throughout, and corrupting the whole bench costs at least $nc+\delta$ once a dissent must be fabricated (the senior premium only raises that cost, reinforcing deterrence); that route attacks only if $B > nc+\delta$.

Proposition 7 (The bench is sized to price out capture)**.** *The smallest bench that deters capture is*

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

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

Proof. By Proposition 6 the partial attack nets at most $(1-p)B-(n-1)c$, which is non-positive iff $n \geq 1+(1-p)B/c$. The fabricated-dissent attack corrupts all n judges and stages a dissent, costing $nc+\delta$ for a conviction worth B, and nets $B-nc-\delta$, non-positive iff $n \geq (B-\delta)/c$. Both routes are unprofitable iff $n \geq \max\{1+(1-p)B/c,\ (B-\delta)/c\}$, so the smallest deterring bench is the ceiling of that maximum. Both bounds increase in B and the first decreases in p. □

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

Remark 4 (Interior benches under uncertain stakes). If the captor's stake B is private, drawn from a distribution G, a bench of size n deters all captors with B below a threshold increasing in n, leaving a residual capture probability that falls in n. Trading this against the cost of convening and vetting judges yields an interior optimal bench that rises with the gravity of the cause – the grading as a smooth comparative static rather than a knife-edge.

8 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 a judge *defer* – copy a senior's vote in place of his own signal – with probability $\gamma > 0$ whenever a senior has voted before him. Under *senior-first* voting every junior has heard a senior and may defer; under *junior-first* voting the juniors commit their signals before any senior speaks, and the seniors defer to no one, so every vote is informative.

Proposition 8 (Junior-first and the unanimity rule are complements)**.** *Junior-first voting yields independent votes, so an honest bench convicts unanimously with probability of order p^n, vanishing in n. Senior-first voting induces deference, so honest unanimity occurs with probability bounded below by the deference rate γ, independent of n. Hence junior-first both minimises the commitment cost $\sim \pi \Pr(\textit{honest unanimity})$ of the acquit-at-unanimity rule and maximises its power to detect capture: honest unanimity becomes rare while captured unanimity does not. The order manufactures the independence the unanimity rule audits.*

Proof. Under junior-first voting a junior moves before any senior, so the deference event never triggers and every vote reflects an independent signal; an honest bench is then unanimous only if all n signals agree, with probability p^n under G and $(1-p)^n$ under I, of order $p^n \to 0$. Under senior-first voting each junior observes a senior and copies with probability γ, so conditional on the first senior convicting, every deferring junior convicts too; honest unanimity then occurs with probability at least that of a deference cascade, bounded below by a function of γ that does not vanish in n. 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 5 (Junior-first protects the bench-size deterrent). Under deference a captor need only corrupt the most senior judge, whom the rest copy, so the cost of swaying the bench collapses to c regardless of n and the grading of Proposition 7 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 6 see that he is pivotal. The order thus does triple duty – independence, pivotal information, and the integrity of the bench-size deterrent.

9 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 – a finding Duggan and

Martinelli (2001) carry to a full spectrum of signals, where unanimity alone leaves the probability of error bounded away from zero as the jury grows. 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. Because that common cause is *state-independent* – it loads the convicting consensus whether or not the defendant is guilty – it does more than blunt aggregation: it inverts the posterior at the top, which the symmetric correlation Ladha studies does not. And where Coughlan (2000) defends the unanimity rule by letting jurors communicate before voting, restoring the sincere voting under which unanimity minimises error, we defend a rule that withholds conviction from unanimity for a different reason – not to coordinate honest jurors but to disarm a bench that may not be honest. The two defences are complementary, the one addressing strategic voting and the other procedural capture.

That the right rule turns on the quality of the information is itself a theme of optimal committee design: a demanding supermajority, even unanimity, is optimal only when signals are accurate enough (Persico, 2004) – a logic our graded benches share, scaling the court to

the gravity, and so the stakes, of the cause. And under deliberation most rules collapse to equivalence, the two unanimity rules alone standing apart (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.

10 Conclusion

We have read the Sanhedrin's capital procedure as a design for judging when the judges may not be free. A bench can be captured, and a captured bench manufactures consensus; so a court that treats unanimity as proof convicts exactly when it has been manipulated. The rule that acquits a unanimous bench (Sanhedrin 17a) is the answer: the one verdict capture can force is the one the court will not honour. From the same primitive follow the pro-acquittal asymmetry, the junior judges who vote first, and the bench that grows with the gravity of the charge – each a part of one defence against a court that may not be its own.

This is a small step, and a first-shot. It holds a single latent probability – that independence has failed – against a single procedure, and leaves to its companions the wider question of how often a rule the law fixed against revision encodes a commitment a theory would later name. What the reading offers is a way to see a perverse rule as a precaution: consensus treated not as the summit of proof but as the place a captured court betrays itself.

References

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

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

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

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

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

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

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

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

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

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

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

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

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.

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.

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.