Review of Economic Studies (2018) **85**, 1897–1935 doi:10.1093/restud/rdx073

Advance access publication 7 December 2017

A Conversational War of Attrition

MORITZ MEYER-TER-VEHN
UCLA
LONES SMITH
Wisconsin
and
KATALIN BOGNAR
Private Sector

First version received October 2013*; Editorial decision July* 2017*; Accepted December* 2017 (*Eds.*)

We explore costly deliberation by two differentially informed and possibly biased jurors: A hawk Lones and a dove Moritz alternately insist on a verdict until one concedes. Debate assumes one of two genres, depending on bias: A juror, say Lones, is *intransigent* if he wishes to prevail and reach a conviction for any type of Moritz next to concede. In contrast, Lones is *ambivalent* if he wants the strongest conceding types of Moritz to push for acquittal. Both jurors are ambivalent with small bias or high delay costs. As Lones grows more hawkish, he argues more forcefully for convictions, mitigating wrongful acquittals. If dovish Moritz is intransigent, then he softens (strategic substitutes), leading to more wrongful convictions. Ambivalent debate is new, and yields a novel dynamic benefit of increased polarization. For if Moritz is ambivalent, then he toughens (strategic complements), and so, surprisingly, a more hawkish Lones leads to fewer wrongful acquittals and convictions. So more polarized but balanced debate can improve communication, unlike in static cheap talk. We also show that patient and not too biased jurors vote against their posteriors near the end of the debate, optimally playing *devil's advocate*. We shed light on the adversarial legal system, peremptory challenges, and cloture rules.

Key words: Cheap talk, Committee decision making, Pivot voting, Juries, Adversarial system, Peremptory challenges, Debate, Deliberation, Polarization, Devil's advocate, Monotone comparative statics, War of attrition, Non-linear difference equations

JEL Codes: D71, D72, D82, D83, C62, C72

1. INTRODUCTION

> It's not easy to raise my hand and send a boy off to die without talking about it first…We're talking about somebody's life here. We can't decide in five minutes. Supposin' we're wrong.
>
> Juror #8 (Henry Fonda), *Twelve Angry Men*

The editor in charge of this paper was Marco Ottaviani.

Economics is not in the business of disputing tastes. And few topics pique an economist so much as seeing how preference diversity is resolved. We have in mind juries, tenure cases, FDA panels, Federal Open Market Committee meetings, etc. In each case, partially informed individuals share an imprimatur to dispassionately arrive at the truth. Two key features of the meetings are that (1) the search for truth falls short of certainty since the debate is a costly endeavour for all involved, and (2) the debaters might disagree on the costs of different mistaken decisions. This article develops a new model of debate in which both the length of debate and the wisdom of its decision reflect the debaters' biases, delay costs, and quality of information.

In the cases we envision, the Bayesian parable of information misses its complexity and intrinsic detail. Debaters cannot simply summarize their insights in one likelihood ratio, and not surprisingly, we do not see this happen. In the movie "Twelve Angry Men", *e.g.*, each juror knew different aspects of the witness testimony; on an FDA panel, each member may specialize in different technical aspects of a proposed new drug. Intuitively, debaters each possess a myriad of "pieces of a puzzle". Our debaters are also duty-bound to arrive faithfully at a decision, and so any argument must be verifiable; but this in practice forces everyone to explain their logic, and carefully adduce all facts: A federal juror must "solemnly swear" to ensure "a true deliverance … according to the evidence". Naturally, it takes more time to explain it at a finer grain. In this story, even debaters with identical preferences take time to distill information to their peers.

To capture all these features of debate, we explore a simple *as if parable*. In lieu of a complex signal space and boundedly rational debators, we substitute coarse communication in a standard Bayesian model, and assume that delay is explicitly costly. We specifically explore the dynamics of costly deliberation by two jurors who must agree upon a conviction or acquittal verdict. We assume that any open non-committal communication has already passed, and instead focus on the *dispositive* communication phase, where conversation "gets real". In this voting parable, each juror incurs a delay cost any period a vote is cast. Two concurring votes ("moved" and "seconded") irreversibly seal the verdict; otherwise, voting continues.

We begin with two biased and partially informed jurors, Lones and Moritz. What emerges is an incomplete information war of attrition where the jurors alternatively argue for their *natural verdict*, conviction for the hawk Lones and acquittal for the dove Moritz, until one concedes.

The coarse communication we study captures the spirit of our as-if parable, veiling the precise jurors' types. For an equilibrium is a sequence of type intervals for each player (Theorem 1). A juror opposes his peer until his threshold surpasses his type, and then concedes. An equilibrium describes a zick-zack threshold path through the type space. Reminiscent of partial equilibrium analysis, triangular deadweight loss deviations from the diagonal measure the decision error costs—an *error of impunity* (wrongful acquittal) or *miscarriage of justice* (wrongful conviction). We then characterize all sequential equilibria in which jurors sincerely vote for their desired verdict. Sincere equilibria are indexed by their "drop-dead" dates (Theorems 2 and 3). In a *deferential equilibrium*, one player concedes by a finite date. Here, arguments end either by an equilibrium protocol or fixed cloture rule—such as in parliamentary debates. But our focal equilibrium has no certain last period. This *communicative equilibrium* intuitively corresponds to the finest grain parsing of the "complex signals" in our motivational story. It is the only stable equilibrium (Theorem 4) when jurors are equally patient and not too biased. Intuitively, deferring is not forwardly rational, since types who deviate can convey their powerful private signals.

We next introduce a fundamental taxonomy of debating genres. When Moritz insists on his acquittal verdict, he incurs explicit delay costs. In a standard, private-value war of attrition, such costs are balanced by the strategic gains of outlasting his rival. Moritz' incentives might well have this flavour: He might prefer to outlast every type of Lones planning to concede next period; for instance, this arises when Lones is very hawkish, and so pushes hard for conviction. Since these incentives are adversarial, we call Moritz *intransigent*. In this case, when Moritz quits, he thinks

the verdict is wrong, but simply throws in the towel. But with less *polarized* debate—when Lones is not so biased—he softens his stance, and Moritz' incentives fundamentally switch. We call him *ambivalent* if he actually prefers to lose out to the strongest types of Lones who concedes next period. This form of debate is more constructive and focused on learning—for if Moritz quits, then he agrees with the verdict, and the conversation secures a meeting of the minds. The debate genre may change over time, and may differ across jurors.

The strategic structure of debate depends on the genre. Strategies in the standard war of attrition are *strategic substitutes*—when one player concedes more slowly, his rival gains less from holding out, and so concedes faster. This well describes equilibrium incentives here with intransigent jurors. But with ambivalent jurors, ours is instead a game of *strategic complements*, in which doggedness begets doggedness: For as Moritz grows more partisan, conceding more slowly, Lones learns less from each delay, and fewer of his types concede. Consistent with this, Proposition 1 finds that debate is always ambivalent when jurors are not too biased. Conversely, Proposition 3 shows that communicative debate by sufficiently patient jurors quickly settles into intransigence. Intransigence also obtains in the continuous time limit of our game.

For sharper predictions about the impact of juror bias or waiting costs, Propositions 4–7 restrict to low juror biases or assume communicative debate has gone on for a while. In these cases, the debate lengthens as bias grows or waiting costs fall. To see their impact on decision errors, assume that jurors are not too biased, so that debate is ambivalent. In this case, a more hawkish Lones pushes harder for convictions, and thereby reduces errors of impunity. Less obviously, his tougher stance elicits so much pushback from the dove Moritz that the *chance of a miscarriage of justice also falls.* Here we see the impact of strategic complements with ambivalence: For Moritz can afford to worry less about errors of impunity when Lones grows more assertive, and so can push more for acquittal. But reflecting our strategic dichotomy, at some point, the tables turn: If jurors grow too biased or patient, debate becomes intransigent. Actions then become strategic substitutes, and so Moritz softens when Lones grows tougher. In this case, a more hawkish Lones still limits errors of impunity, but also leads to more miscarriages of justice; however, if jurors are symmetric and *both* grow more biased, then both decision errors fall. This theoretical insight has applied implications—*e.g.* it intimates as to why one might wish to limit the number of peremptory juror challenges, for *a more balanced and polarized jury best determines the truth*. An advantage of our model is that it is identifiable: Since our predictions vary in the bias and cost parameters, one can identify them from observables.

Our as-if model yields a key intuitive feature of debate: Jurors may eventually play *devil's advocate*: Lones might well acquit if he could decide the verdict unilaterally as a dictator, and yet persist in voting to convict. As Lones the debater pays a deliberation cost, one might think him more eager to concede than the dictator, who can end the debate; however, seconding a proposal ends the game, whereas holding out retains the option value of conceding later in light of new information about Moritz's type. Option value is an important element of dynamic debate, and offers a key contrast with static committee models. We prove in Proposition 8 that devil's advocacy always arises, as long as jurors are not too biased and delay is not too costly.

Our article is technically innovative in many ways. Equilibria are characterized by a possibly infinite sequence of thresholds obeying a non-linear second-order difference equation. Since we know of no general method of solving such a dynamical system, we develop new methods to establish existence and uniqueness. Our existence theorem exploits the Jordan curve theorem as a generalization of the intermediate value theorem. Our uniqueness theorem and comparative statics critically exploit an assumption that jurors' types have a log-concave density. Many signal distributions obey this condition, adapted from Smith *et al*. (2016). This condition disciplines the best response functions, since each updates from a truncated signal. We also recursively apply

monotone methods to show that the dynamical system is saddle point stable, and our equilibrium loosely resembles a balanced growth path, familiar to macroeconomists.

Related Literatures. The cheap talk literature started by Crawford and Sobel (1982) also explores communication by informed individuals before an action. Our model specifically assumes that jurors must agree on a verdict—a motion "moved" and then "seconded" seals the verdict. Another key difference is that our communication is not free. The possibility of trading off the chance to achieve one's favourite verdict for the extra delay costs overturns a key insight of the cheap talk literature: namely, greater bias leads to worse decisions. There, greater bias renders communication less transparent, and thereby inflates decision errors. In stark contrast, we find that slightly partisan jurors arrive at the truth more often than unbiased jurors.

As seen in the dynamic cheap talk literature, Forges (1990), Aumann and Hart (2003), and Krishna and Morgan (2004), dynamic communication can convey more information than static communication. As seen in Goltsman *et al*. (2009), dynamic communication formally allows an informed agent to credibly commit to send a mixed signal. In contrast, we explicitly model the optimal level of communication when any communications are costly; further, the willingness to persevere is a credible signal of the strength of one's signal.

The committee decision literature, surveyed in Li and Suen (2009), explores free information transmission before a vote.[1] Adding some structure, Li *et al*. (2001) (LRS) allow jurors to cast multiple votes. Our model with vanishing delay costs approximates LRS, but our focal communicative equilibrium has no counter-part in their model. The limit with vanishing delay costs differs from zero delay costs, for costs can be amplified endogenously in equilibrium.

Another strand of the committee literature focuses instead on *public information acquisition*. In this research thread, Chan *et al*. (forthcoming) [CLSY] is the closest work to us: They consider a war of attrition by voters with ordinally aligned preferences who observe a continuous time public information process.[2] In contrast, our jurors have *already* witnessed the trial evidence, our panel has researched a shuttle explosion, or our committee has seen and read an assistant professor's research. For instance, FDA panels do not convene until Phase 1–3 trials have ended, and a new drug application is received. We analyse the subsequent deliberation when debaters are thus endowed with their *private signals*, and seek to learn about each other's signals. CLSY explore the majority requirements for the vote—a moot point for our jury of two. With unanimity, their war of attrition analysis is akin to our intransigent debate: behaviour by the pivotal voters exhibits strategic substitutes, since a tougher stance by the most hawkish juror induces the most dovish juror to soften. In contrast, our new ambivalent debating genre exhibits strategic complements, and has no counterpart in their analysis.[3,4]

Our article contributes to the war of attrition literature. For instance, Gul and Pesendorfer (2012) consider a complete information war of attrition played by two parties with ordinally opposed preferences, and so each seeking to win. In contrast, we arrive at a meeting of the minds with ambivalent debate—when a juror concedes, he is genuinely convinced of his peer's perspective. In a companion paper Meyer-ter-Vehn *et al.* (2017), with unbiased and equally patient jurors, any cloture rule that truncates debate at a fixed date lowers welfare. In contrast, asymmetric

1. See Coughlan (2000), Piketty (2000), Austen-Smith and Feddersen (2006), Gerardi and Yariv (2007).

2. With common preferences, information is a public good; Persico (2003), Gerardi and Yariv (2008), Gershkov and Szentes (2009) study how to incentivize committee members to provide this public good.

3. Sometimes with majority rule, strategic complements can also arise in CLSY when biased factions of jurors vie for the vote of an impatient swing voter, by relaxing standards. But in this case, greater bias leads jurors to *soften their stance*, which thereby *magnifies* decision errors—the opposite to our earlier takeout message.

4. The strategic exercise in CLSY is also formally static, with every player optimizing at time-0 over a quitting time, as if in an n-player auction. Ours exploits sequential equilibrium refinements, using off-path inferences. Moreover, no player could at time-0 foresee his stopping time had we assumed three or more players.

equilibria in which one juror concedes immediately to his insistent peer, or deadlines that enforce such early agreements, generally increase efficiency when there is a conflict of interest, as in Gul and Lundholm (1995) or Damiano *et al*. (2012).[5]

We next introduce and analyse the model, highlighting its novel Bayesian aspects. A two period example gives a foretaste of our results. We prove most results in the Appendix.

2. THE MODEL

2.1. *The extensive form game*

Two jurors $i=L,M$, Lones and Moritz, alternately propose in periods $t=0,1,2,\ldots$ to *convict* or *acquit*, $\mathcal{C}$ or $\mathcal{A}$, a defendant of a crime. Lones proposes a verdict in period zero. Moritz replies in period one with his own proposal. The game ends if he agrees; otherwise, Lones responds in period two with a proposed verdict, and so on. The game ends when two consecutive verdicts concur—namely, a unanimity rule.

A priori, the defendant is equilikely to be *guilty* or *innocent*—states $\theta=\mathcal{G},\mathcal{I}$. Jurors are partially informed: each has privately observed a signal about the defendant's guilt. The conditionally iid signals $\lambda,\mu\in(0,1)$ are *private beliefs* that the defendant is guilty, *i.e.* $\theta=\mathcal{G}$.

Lones is a hawk, hurt weakly more by errors of impunity, and Moritz a dove, hurt more by miscarriages of justice. Jurors' *decision costs* are $1+\beta_L, 1-\beta_M$ for an *actual error of impunity*, *i.e.* acquitting the guilty, and $1-\beta_L, 1+\beta_M$ for an *actual miscarriage of justice*, *i.e.* convicting the innocent, where $\beta_i\geq 0$ is the *bias* of juror i. Jurors share the same cardinal preferences over verdicts if $\beta_L=\beta_M=0$. We assume $\beta_L,\beta_M<1$, ensuring identical ordinal preferences: convict the guilty and acquit the innocent, precluding partisans, who always wish to convict or acquit.[6]

That $0<\lambda,\mu<1$ and $\beta_L,\beta_M<1$ reflects standard jury instructions to "be open-minded", for in this case, jurors are willing to change their mind given enough evidence. Circuit court jurors are advised: "While you're discussing the case, don't hesitate to reexamine your own opinion and change your mind if you become convinced that you were wrong" (Ed Carnes, 2016).

The jurors find debate time costly: Juror i incurs a *waiting cost* $\kappa_i>0$ per delay period. To avoid trivialities, we assume $\kappa_i<1-\beta_i$.[7] We say that *jurors are symmetric* if $\kappa_L=\kappa_M$ and $\beta_L=\beta_M$. Jurors minimize losses, namely, the expected sum of waiting costs and decision costs. So they are risk-neutral and do not discount future payoffs.[8]

All told, the game resembles a war of attrition: a stopping game in which each juror trades off the exogenous cost of continuing against the strategic incentives to insist on his preferred verdict; however, this preferred verdict may change as he learns his peer's type.

2.2. *Transforming signals*

We represent signals as log-likelihood ratios, with different reference states.[9] Jurors' transformed *types* are $\ell=\log(\lambda/(1-\lambda)), m=\log((1-\mu)/\mu)$. No signal is perfectly revealing, and the *common unconditional type density* f is positive and symmetric $f(x)\equiv f(-x)$ of $x=\ell,m$, with cdf F. Since

5. That paper can be viewed as a variant of our article with binary, perfectly informative signals.
6. Our analysis can be used for this case, but it presents additional technical hurdles, and so we avoid it.
7. This bound ensures that juror i, say Moritz, prefers a conviction in $t+1$ over an acquittal in t when the defendant is guilty. Without this assumption, Nixon-China debates terminate immediately.
8. Debates typically last hours, days, weeks, or maybe months, where discounting is not important.
9. For simplicity, we denote random variables and their realizations by the same notation. We flag randomness whenever it might be in doubt. We justify in Section A.1 that we may first specify an unconditional density f.

the random signals λ, μ are conditionally independent, so too are the transformed (random) types ℓ, m.

Let $p(\ell, m)$ be the conditional probability of guilt given types ℓ, m. Bayes' rule implies

$$p(\ell, m) = \frac{e^{\ell - m}}{e^{\ell - m} + 1}. \tag{1}$$

Unlike an actual error of impunity, an *error of impunity* is the ex post event of acquittal despite $p(\ell, m) > \frac{1}{2}$.[10] A *miscarriage of justice* likewise means that conviction occurs despite $p(\ell, m) < \frac{1}{2}$.

Any type y juror entertains the conditional probability density $f(x|y)$ that his colleague's type is x. Let $h(x, y)$ be the unconditional joint density and $r(x, y) \equiv h(x, y)/(f(x)f(y))$ the *correlation factor*. This yields the conditional density that one's peer is type x, given the realized type y, updating from the common type density: $f(x|y) = f(x)r(x, y)$. In Section A.1, we show that $r(x, y) = 2(e^x + e^y)/((1 + e^x)(1 + e^y))$, and that $r(x, y)$ and $h(x, y)$ are log-submodular. For intuitively, since signals about the state are affiliated,[11] so too are their log-likelihood ratios; but then the inversely defined random types ℓ, m are *negatively* affiliated, as Figure 1a highlights.

2.3. *Strategies and payoffs*

Lones' initial vote fixes the debate roles for the rest of the game. If he initially proposes $\mathcal{C}$, then jurors enter the *natural subgame*, in which each argues for his *natural verdict*—$\mathcal{C}$ for Lones and $\mathcal{A}$ for Moritz—until conceding;[12] otherwise, they enter the *Nixon-China* subgame where each argues for his *unnatural verdict* until conceding.[13] Since either subgame is a stopping game, we describe pure strategies by the planned *stopping times*—the first period a juror plans to concede if the game has not yet ended.[14] So Moritz has a strategy described by two (odd) periods in which he first concedes to Lones after either initial proposal, while Lones' strategy consists of his initial proposal and his planned (even) concession period. We say that Moritz *convinces* Lones in period t, say, if Lones quits in period t.

Strategy profiles are *equivalent* if they imply the same outcome; for any juror's strategy, call two strategies of the other juror *equivalent* if the strategy profiles are equivalent. We find Bayes Nash equilibria (BNE), and later prove that any BNE is equivalent to a sequential equilibrium.

3. PRELIMINARY EQUILIBRIUM ANALYSIS

3.1. *Monotonicity*

Lones' initial vote fixes an ordering on types. In the natural subgame, a *stronger* type of Lones and Moritz is higher, and so more convinced that the proposed verdict is right. In the Nixon-China subgame, *stronger* types of Lones and Moritz are lower. In a *monotone* strategy, whenever some juror type holds out until period t, a stronger type surely holds out until then.

10. Here, *ex post* means conditional on both types ℓ, m, but not on the unknowable state $\theta = \mathcal{G}, \mathcal{I}$.

11. Random variables with a (log-submodular) log-supermodular density are *(negatively) affiliated*.

12. Given Nature's initial move, there are no proper subgames. We use this term for the *subform*.

13. The political metaphor "Nixon in China" refers to the idea that "Only a politician or leader with an impeccable reputation of upholding particular political values could perform an action in seeming defiance of them without jeopardizing his support or credibility" (Wikipedia) such as the hawk Lones arguing for acquittal, or the dove Moritz arguing for conviction in our case.

14. A player who stops in period t also plans to stop at all later periods $t' > t$, precluded by his earlier actions. Similarly, Lones plans to stop at any period t after the initial verdict which he does not choose.

Lemma 1. (Single Crossing Property). *If a type prefers to hold out from period t to $t' > t$, then any stronger type prefers to do so, and strictly so if period t is hit with positive probability. Every best response strategy of a juror to any strategy is thus equivalent to a monotone strategy.*

Stronger types hold out longer as they are not only more convinced of their position, but also more sure that their peer entertains a weaker opposing signal, given their negative correlation.

Lemma 1 yields a skimming property of equilibria, familiar in the bargaining literature: Stronger types quit in every period until the end. Moritz' monotone strategy in the natural subgame is described by a weakly increasing sequence of odd-indexed cutoff types $(x_t)_{t\in 2\mathbb{N}+1}$, where x_t is his supremum type that concedes by period t. Lones' monotone strategy in the natural subgame is likewise described by a weakly increasing sequence of even-indexed cutoff types $(x_t)_{t\in 2\mathbb{N}+2}$; monotone strategies in the Nixon-China subgame are described by weakly decreasing sequences $(x_{-t})_{t\in 2\mathbb{N}+1}$ and $(x_{-t})_{t\in 2\mathbb{N}+2}$. In other words, type intervals of Lones and Moritz stop in alternating periods; moreover, negative period indexes simply flag that the threshold corresponds to Nixon-China debate. Hereafter, we assume monotone strategies.

Consider next period zero. Lones' strategy is *sincere* if he proposes to convict when his type indicates guilt, *i.e.* $\ell > x_0$ for some x_0, and acquit otherwise. A sincere strategy is *responsive* if not all types vote for the same verdict, *i.e.* $|x_0| < \infty$.[15] Finally, a strategy of Moritz is *agreeable* if almost all of his types m either plan to second Lones' initial proposal to convict, namely, $m < x_1$, or to second the initial proposal to acquit, *i.e.* $m > x_{-1}$; this is equivalent to $x_{-1} \leq x_1$.

Lemma 2. *Lones' best reply to an agreeable, monotone strategy of Moritz is sincere. Conversely, any best reply of Moritz to a sincere, monotone strategy of Lones is equivalent to an agreeable strategy. So up to equivalence, Lones is sincere in equilibrium if and only if Moritz is agreeable.*

A sincere agreeable responsive equilibrium is characterized by cutoffs $(x_t)_{t\in\mathbb{Z}}$ with $|x_0| < \infty$, and:[16]

$$\begin{gathered} -\infty \leq \cdots \leq x_{-3} \leq x_{-1} \leq x_1 \leq x_3 \leq \cdots \leq \infty \\ -\infty \leq \cdots \leq x_{-4} \leq x_{-2} \leq x_0 \leq x_2 \leq x_4 \leq \cdots \leq \infty \end{gathered} \tag{2}$$

Figure 1b depicts the equilibrium outcomes in this sincerere agreeable case.[17]

3.2. *The propensity to hold out*

We now characterize equilibrium cutoffs (x_t) in terms of indifference conditions. When type y of Lones, say, faces type x of Moritz, conviction is the correct verdict with chance $p(y,x) \equiv 1 - p(x,y)$ (recalling (1)) and securing it avoids an actual error of impunity, thereby lowers decision costs by $1+\beta_L$; acquittal is the correct verdict with chance $p(x,y)$ and securing it avoids an actual miscarriage of justice, and thereby lowers decision costs by $1-\beta_L$. Summing up, the net change in *expected decision cost* from securing Lones' natural conviction verdict is $(1+\beta_L)(1-p(x,y)) - (1-\beta_L)p(x,y) = 1 - 2p(x,y) + \beta_L$; an analogous argument applies for Moritz. Using (1), write this

15. This ordinal sincerity notion is weaker than the cardinal notion in the strategic voting literature, which requires that $x_0 = 0$ when jurors are unbiased.

16. Thresholds need not all alternate; only odds and evens need be sorted. In the two-period equilibrium in Section 4, *e.g.*, if we abandon symmetry and assume that Moritz is a much stronger dove than is Lones a hawk, then $x_0 > x_1$. Still, we draw all plots with cutoffs fully ordered by their indexes.

17. Online Appendix Section B.1 explores non-sincere and non-agreeable equilibria.

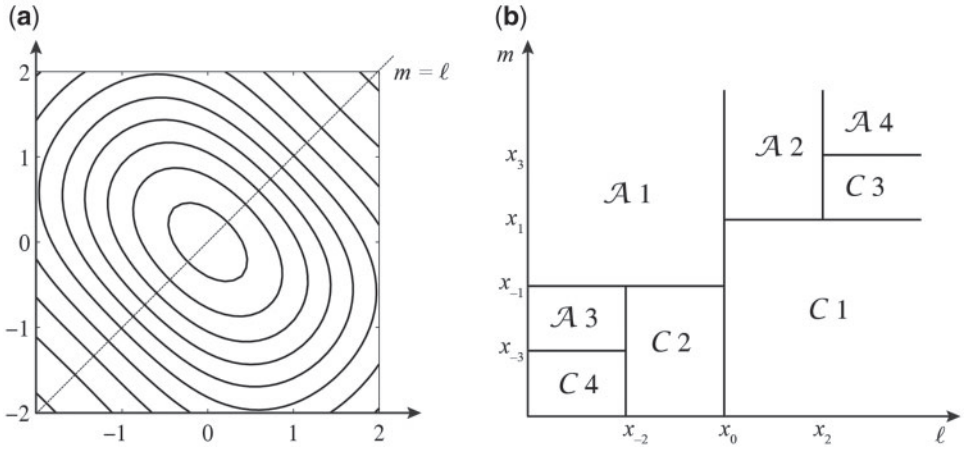

FIGURE 1

Joint signal density and equilibrium outcomes. (a) Plots contour lines of the negatively affiliated joint type density $h(\ell,m)$. (b) Depicts the outcomes of a sincere agreeable monotone strategy profile. Here, $\mathcal{C}1$ means conviction in period 1, etc. Lones' types and (even) cutoffs are on the horizontal axis; Moritz' types and (odd) cutoffs on the vertical axis

expected decision payoff in terms of the *type difference* $\delta \equiv x-y$:

$$\Delta(\delta,\beta_i)=\frac{1-e^{\delta}}{1+e^{\delta}}+\beta_i. \tag{3}$$

As seen in Figure 2a, this payoff falls in the type difference δ, since a stronger peer type x lowers the chance that the natural verdict is correct.

Jurors solve an infinite horizon stopping problem, but it suffices to plan for two periods. Denote the conditional density $f(x|y,x\geq\underline{x})=f(x|y)/[1-F(\underline{x}|y)]$. If juror i of type y believes that his peer's types $x<\underline{x}$ have conceded, and that those in $[\underline{x},\bar{x}]$ will next do so, then his expected payoff gain in the natural subgame from holding out more period is the *propensity function*:

$$\Pi_i(\underline{x},y,\bar{x})\equiv\int_{\underline{x}}^{\bar{x}}(\Delta(x-y,\beta_i)-\kappa_i)f(x|y,x\geq\underline{x})dx-\int_{\bar{x}}^{\infty}2\kappa_i f(x|y,x\geq\underline{x})dx. \tag{4}$$

The integrand is the *net decision payoff*, *i.e.* first $\Delta(x-y,\beta_i)-\kappa_i$ for $x\in(\underline{x},\bar{x})$ and then jumping to $-2\kappa_i$ for $x\geq\bar{x}$. It measures the net benefits of an immediate concession in the next period, conditional on the peer type x. It includes a decision payoff gain for *weak* conceding peer types x with $\Delta(x-y,\beta_i)>0$, and a decision payoff loss for *strong* conceding peer types x with $\Delta(x-y,\beta_i)<0$—both net of one period delay costs of κ_i. The two period delay cost $2\kappa_i$ is incurred if the peer does not concede.

Similarly, for the Nixon-China subgame, define the propensity to hold out $\hat{\Pi}_i(\underline{x},y,\bar{x})$ by juror i of type y when types $x>\bar{x}$ have already conceded and types $x\in[\underline{x},\bar{x}]$ will next concede:

$$\hat{\Pi}_i(\underline{x},y,\bar{x})\equiv\int_{\underline{x}}^{\bar{x}}(-\Delta(x-y,\beta_i)-\kappa_i)f(x|y,x\leq\bar{x})dx-\int_{-\infty}^{\underline{x}}2\kappa_i f(x|y,x\leq\bar{x})dx. \tag{5}$$

In contrast to (4), the expected decision payoff now pertains to securing one's *unnatural* verdict, and hence has the opposite sign, $-\Delta(x-y,\beta_i)$. Indifference by cutoff types x_t and x_{-t} requires:

$$\Pi_{i(t)}(x_{t-1},x_t,x_{t+1})=0=\hat{\Pi}_{i(t)}(x_{-(t+1)},x_{-t},x_{-(t-1)}) \tag{6}$$

for $t=1,2,\ldots$, where $i(t)=L$ for even periods t, and $i(t)=M$ for odd t. This is the discrete first-order condition between periods $t-1$ and $t+1$ for the (omitted) infinite horizon Bellman value.

When Lones employs an initial cutoff type x_0 in a sincere agreeable responsive strategy profile, and plans to concede in period two, should Moritz hold out in period one, using his next cutoffs x_1 and x_{-1}, then Lones' initial propensity to convict equals:

$$\bar{\Pi}_L(x_{-1},\ell,x_1)\equiv\int_{-\infty}^{x_{-1}}\kappa_L f(m|\ell)dm+\int_{x_{-1}}^{x_1}\Delta(m-\ell,\beta_L)f(m|\ell)dm-\int_{x_1}^{\infty}\kappa_L f(m|\ell)dm. \quad (7)$$

Indeed, any type $m\leq x_{-1}$ of Moritz immediately agrees in the natural subgame, but holds out in period one of the Nixon-China subgame. Both initial proposals lead to a conviction, but proposing to convict reduces Lones' waiting costs by κ_L. Types $m\in(x_{-1},x_1)$ of Moritz agree at once in both subgames, whereupon Lones' proposal fixes the verdict. In this case, Lones' decision payoff from proposing to convict equals $\Delta(m-\ell,\beta_L)$. Finally, types $m\geq x_1$ of Moritz agree immediately in the Nixon-China subgame, but hold out in period one of the natural subgame. Hence, both initial proposals lead to the same acquittal verdict, but proposing to convict raises waiting costs by κ_L. Lones' type $m=x_0$ is indifferent between initially voting convict or acquit (and conceding immediately if Moritz does not agree) if:

$$\bar{\Pi}_L(x_{-1},x_0,x_1)=0. \quad (8)$$

We next show that indifference conditions (6) and (8) characterize equilibrium. Cutoffs are *tight* if whenever all types of one juror concede, all remaining types of the other juror thereafter hold out forever, *i.e.* if $x_t=\infty$ at some odd period t, say, then $x_{t'}=x_{t-1}$ for even $t'>t$.

Theorem 1. (Characterization). *A sincere agreeable responsive equilibrium is equivalent to tight cutoffs* (x_t) *that obey monotonicity (2), indifference (6), and (8) if finite, and* $|x_0|<\infty$. *Conversely, any such cutoffs define a sincere agreeable responsive equilibrium.*

3.3. *Equilibrium existence*

For insight into equilibria, assume that Lones insists on his initial vote—acquit for $\ell<x_0$ and convict for $\ell>x_0$—forever after. Since Moritz cannot affect the verdict, he defers at once. This strategy profile is a Bayes-Nash equilibrium for suitable x_0. It corresponds to an asymmetric outcome of a standard, private value war of attrition, *e.g.* Riley (1980). But deference can arise in any period t in our game. Had we assumed private juror values (over verdicts), a strategy profile in which, say, Lones surely concedes in period t unravels: For (1) no type of Moritz concedes in period $t-1$, and so (2) all remaining types of Lones concede in period $t-2$, and so on. But in our common values setting, step (1) of this unraveling logic breaks down: weak types of Moritz do not want win the debate against the remaining strong types of Lones in period $t-1$ and hence concede; this in turn gives Lones an incentive to hold out in period $t-2$.

A (σ,τ)-*equilibrium* is a minimal pair of *drop-dead dates* $1\leq\sigma,\tau\leq\infty$ such that debate ends by period σ of the Nixon-China subgame and τ of the natural subgame. This equilibrium is *deferential in the Nixon-China subgame* if $\sigma<\infty$, and otherwise communicative. It is *deferential in the natural subgame* if $\tau<\infty$, and otherwise *communicative*.[18] It is deferential if $\sigma,\tau<\infty$, and *communicative* if $\sigma=\tau=\infty$.

18. One might think of "deferential" equilibria instead as "insistent" equilibria—namely, all types of Moritz, say, dig in their heels and insist on conviction. But Lones optimally defers in period t as long as few enough of Moritz' types plan to concede in $t+1$. We therefore focus on Lones' deference, rather than Moritz' insistence.

Theorem 2. (Existence). *A (σ,τ)-equilibrium exists for all integers (σ,τ), with $1\leq\sigma,\tau\leq\infty$.*

The drop-dead dates of a deferential equilibrium are enforced strategically in our open-ended game. But one can also view it as the longest possible equilibrium in a truncated game where drop-dead dates are enforced by protocol or regulation, such as cloture rules in a parliament.

By Theorem 1, equilibrium cutoff vectors (x_t) are described by a second-order difference equation, solving (6) and (8); deferential equilibria also obey the boundary conditions $x_{-\sigma}=-\infty$, $x_\tau=\infty$, and the communicative equilibrium obeys transversality conditions. But there is no general existence or uniqueness methodology for non-linear second-order difference equations. Our existence proof in Section A.5 is intrinsically topological, while our uniqueness proof for small bias, namely, the argument for Theorem 5 in Section A.10, uses monotonicity methods.

3.4. *Equilibrium stability*

We next ask which equilibria in Theorem 2 obey stronger and more robust solution concepts.

Theorem 3. (Sequentiality). *The communicative equilibrium is a sequential equilibrium. Any deferential equilibrium is a sequential equilibrium.*

Proof. In a communicative equilibrium, all information sets are reached on path, and so Bayes' rule determines beliefs; the resulting assessment therefore constitutes a sequential equilibrium.

Next consider a deferential equilibrium, say with Lones conceding in periods $\sigma,\tau<\infty$. If he unexpectedly holds out in period τ, then any beliefs over his random type ℓ derive from some sequence of completely mixed strategies in a sequential equilibrium; that is, the consistency requirement has no bite. For Moritz may interpret the failure to concede as a tremble of weak types; formally, his type m may believe that Lones' type ℓ obeys $\Delta(\ell-m)>\kappa_M$ almost surely. With such beliefs, and expecting that Lones is about to concede, Moritz wishes to hold out. ‖

Yet deferential equilibria do represent a communication failure. Lones concedes not because he is convinced that Moritz is right, but because Moritz refuses to concede. This could not happen if Moritz had to interpret off-path behaviour by Lones as a signal of strength, rather than as a mistake. Inspired by a definition for finite games in Cho (1987), we say that a sequential equilibrium obeys *forward induction* if either juror who observes a deviation from the equilibrium path must assign probability zero to any types of his peer for whom the observed deviation is not sequentially rational for equilibrium beliefs and *any* conjecture about future play.

Theorem 4. (Stability). *The communicative equilibrium satisfies forward induction. If jurors are equally patient, then (1) equilibria that are deferential in the natural subgame violate forward induction if and only if the biases $\beta_L,\beta_M\geq 0$ are both sufficiently small; and (2) equilibria that are deferential in the Nixon-China subgame violate forward induction.*

Forward induction has no bite in a communicative equilibrium, as every information set is hit on the equilibrium path. Next consider a deferential equilibrium, say, with Lones conceding in period τ of the natural subgame. If he unexpectedly holds out, then Moritz must blame this deviation either on Lones' bias or his information. If Lones is very biased, then Moritz need not infer that Lones holds compelling information for guilt, and Moritz may insist on acquitting. But if Lones is not too biased, then only strong types profit from holding out; forward induction obliges Moritz to acknowledge Lones' strong information. Still, we can rationalize Moritz' insistence on acquitting if Moritz is very biased. But otherwise, his weakest remaining type is sufficiently

convinced of guilt that he concedes, and the equilibrium unravels. All told, forward induction prunes deferential equilibria when neither juror is too biased.[19,20]

Stability prunes deferential equilibria in the Nixon-China subgame: Since jurors argue against their bias, suddenly holding out for the unnatural verdict betrays strong information.

Theorem 4 selects the communicative equilibrium when jurors are not too biased. This choice also follows from reputational concerns by the logic of Abreu and Gul (2000). For assume that not only can either juror be a rational type ℓ, but with a small chance, he is a "behavioural type" who adamantly never changes his verdict. This prunes any deferential equilibria with early drop-dead dates. Indeed, if Moritz expects all rational types of Lones to concede by period τ, and Lones instead holds out at period τ, then Moritz in period $\tau+1$ infers that Lones is the behavioural type; Moritz then concedes at once, undermining Lones' deference in period τ.[21]

Besides failing stability, deferential equilibria are actively discouraged in legal settings. Standard juror instructions remind them of their *duty to deliberate*—for instance: "While you're discussing the case, … don't give up your honest beliefs just because others think differently or because you simply want to get the case over with" (Ed Carnes, 2016).

3.5. *Preliminary analysis for the characterization results*

3.5.1. Distributional assumptions. Subsequent results also require a type density restriction:

($\star$) The type density f is log-concave and has a bounded hazard rate.

Log-concavity is satisfied by many standard distributions, and implies a monotone hazard rate $f/(1-F)$. A bounded hazard rate is less standard, but holds for the logistic and Laplace distributions.[22] Since the hazard rate is monotone and bounded, it finitely converges, say to $\gamma^{-1}<\infty$. The inverse γ of this tail hazard rate measures the thickness of the tail of the jurors' type distribution, and intuitively is a measure of signal *informativeness*.

3.5.2. Propensity function properties. When peer types $x<\underline{x}$ of a juror y have conceded, and types in $[\underline{x},\bar{x}]$ next concede, we call the *lower gap* $\underline{\delta}\equiv y-\underline{x}$ and the *upper gap* $\bar{\delta}\equiv\bar{x}-y$. Analogous to (4), we define a *gap propensity* function $\pi_i(\underline{\delta},y,\bar{\delta})\equiv\Pi_i(y-\underline{\delta},y,y+\bar{\delta})$, obeying:

$$\pi_i(\underline{\delta},y,\bar{\delta})=\int_{-\underline{\delta}}^{\bar{\delta}}(\Delta(\delta,\beta_i)-\kappa_i)f(y+\delta|y,\delta\geq-\underline{\delta})d\delta-\int_{\bar{\delta}}^{\infty}2\kappa_i f(y+\delta|y,\delta\geq-\underline{\delta})d\delta. \tag{9}$$

19. This logic proves the instability of deferential equilibria if Lones' deference in period τ is rationalized by all remaining types of Moritz insisting on acquit in period $\tau+1$. We extend this argument in Section A.6 to the case with a non-empty set of Moritz' conceding types, but so small that no type of Lones wishes to hold out in period τ.

20. This argument suggests that deferential equilibria can be stable when at least one juror is very biased. Indeed, the proof in Section A.6 contains the key arguments to construct such equilibria.

21. The last step of this argument assumes that Lones puts small probability on Moritz' "behavioural type", conditional on reaching period τ. For a small ex-ante chance of behavioural types, this condition holds for small τ, but eventually fails. Thus, any positive chance of behavioural types rules out not just short deferential equilibria, but also the communicative equilibrium, since jurors eventually grow convinced they are facing a behavioural peer, and then concede. All told, we select long deferential equilibria. We conjecture that the communicative equilibrium emerges as the limit of the long deferential equilibria as the chance of behavioural types vanishes.

22. The normal does not have this property. But distributional assumptions are usually imposed on the density of private beliefs $\phi(\lambda)$, and not the density of log-likelihood ratios $f(x)$. Lemma A.1 shows that the hazard rate of f is bounded if $\lim_{\lambda\to 0}\phi'(\lambda)\lambda/\phi(\lambda)$ finitely exists—as with all Beta-distributions, including the uniform.

We now partition the gaps into intervals with endpoints $\underline{b}_i < b_i < \bar{b}_i$, given the bias β_i. These are the respective roots of the net decision payoff and decision payoff, $\Delta(\underline{b}_i, \beta_i) - \kappa_i = 0 = \Delta(b_i, \beta_i)$, and the crossing point of the integrands in the two integrals of (9), or, $\Delta(\bar{b}_i, \beta_i) = -\kappa_i$:

$$\underline{b}_i \equiv \log \frac{1+\beta_i-\kappa_i}{1-\beta_i+\kappa_i} \qquad \text{and} \qquad b_i \equiv \log \frac{1+\beta_i}{1-\beta_i} \qquad \text{and} \qquad \bar{b}_i \equiv \log \frac{1+\beta_i+\kappa_i}{1-\beta_i-\kappa_i}. \tag{10}$$

In other words, juror i is indifferent between verdicts if his peer's type exceeds his own by b_i; he is willing to wait an extra period to achieve his natural verdict if this type difference is $\underline{b}_i$; and he is willing to wait an extra period to get his unnatural verdict if the type difference is $\bar{b}_i$.

($P1$) *The gap propensity* π_i *quasi-increases*[23] *in the lower gap* $\underline{\delta}$*, and is negative for* $\underline{\delta} < -\underline{b}_i$.
($P2$) *The gap propensity* π_i *quasi-increases in the type* y;
($P3$) *The gap propensity* π_i *is hump-shaped in the upper gap* $\bar{\delta}$*, with maximum at* $\bar{\delta} = \bar{b}_i$.
($P4$) *The gap propensity* π_i *increases in the bias* β_i*, and decreases in the waiting cost* κ_i.

To see why property ($P1$) holds, consider Figure 2a. For larger gaps $\underline{\delta}$ with $-\underline{\delta} < \underline{b}_i$, the gap propensity π_i grows in $\underline{\delta}$; intuitively, as $\underline{\delta}$ rises, more weak peer types concede, and the willingness to concede rises. For all smaller lower gaps $\underline{\delta}$ with $-\underline{\delta} > \underline{b}_i$ (not pictured), the gap propensity $\pi_i < 0$. So π_i either increases or is negative, *i.e.* it quasi-increases (proof in Section A.4).

Property ($P2$) intuitively follows because a higher type shifts probability of the negatively affiliated peer's type left towards weaker types and the positive area DG (proved in Section A.4).

Next, the proof of property ($P3$) considers three cases, with the last one pictured in Figure 2a. For upper gaps $\bar{\delta} < \underline{b}_i$, the positive area decision gain DG rises in $\bar{\delta}$. For upper gaps $\bar{\delta} \in [\underline{b}_i, \bar{b}_i]$, the negative decision loss area DL falls in $\bar{\delta}$, and so the propensity rises. Finally, for upper gaps $\bar{\delta} > \bar{b}_i$, the negative area DL rises in $\bar{\delta}$. This proves ($P3$).

Property ($P4$) simply follows from (9).

Our article sometimes focuses on the debate when types are large. Here, we have a diagonal monotonicity property with a directional derivative flavour:

($P5$) *For large enough* y*, there is* $\varepsilon > 0$ *with* $\partial \pi_i / \partial \underline{\delta} > (1+\varepsilon) |\partial \pi_i / \partial \bar{\delta}|$ *when* $\pi_i(\underline{\delta}, y, \bar{\delta}) = 0$.

Proved in Section A.4, this asserts that the propensity is strictly more sensitive to its first than third argument. For these partial derivatives are proportional to the density $f(y+\delta | y, \delta \geq -\underline{\delta})$ at $\delta = -\underline{\delta}, \bar{\delta}$, and this density falls exponentially over the (boundedly positive) interval $[y-\underline{\delta}, y+\bar{\delta}]$.

To analyse the limit game, define the *limit propensity* $\pi_i^{\infty}(\underline{\delta}, \bar{\delta}) \equiv \lim_{y \to \infty} \pi_i(\underline{\delta}, y, \bar{\delta})$. We show in Section A.4 that π_i^{∞} exists and inherits the derivative properties ($P1$) and ($P3$)–($P5$) of π_i.

4. TWO PERIOD DEBATE: AN ILLUSTRATIVE EQUILIBRIUM

As a foretaste of our general theory, we explore the deferential equilibrium with drop-dead dates $\sigma = 1$ and $\tau = 2$. Here, only a single rebuttal is possible, in which Moritz may push for acquittal against an initial conviction proposal by Lones. So Lones and then Moritz propose verdicts, and

23. A function $f(x)$ *quasi-increases* if $f(x) \geq 0$ implies $f(x') > 0$ for all $x' > x$ and *quasi-decreases* if $f(x) \leq 0$ implies $f(x') < 0$ for all $x' > x$; equivalently, f is (strictly) single-crossing from below/above. Thus, a smooth function $f(x)$ is quasi-concave iff $f'(x)$ quasi-decreases.

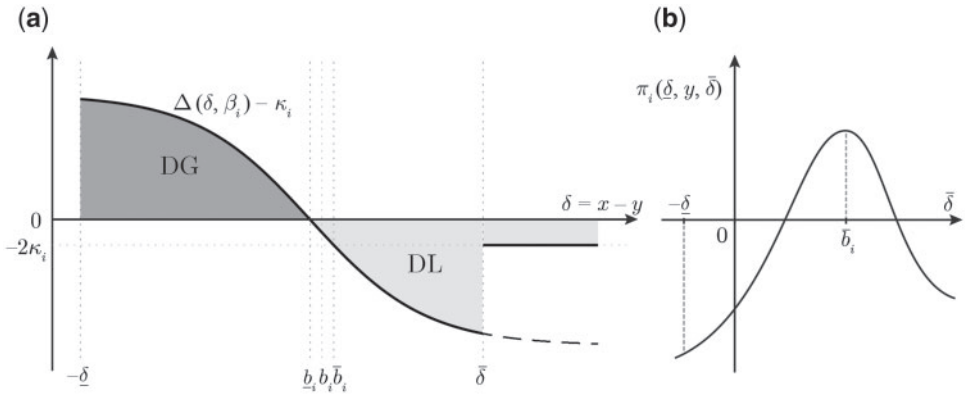

FIGURE 2

The net decision payoff and gap propensity function. Juror i's net decision payoff as a function of the gap δ is positive for $\delta < \underline{b}_i$ and negative for $\delta > \underline{b}_i$ — the gain DG and loss DL; the latter includes the delay costs if the peer holds out. Its propensty integral in panel (b) is hump-shaped in the upper gap $\bar{\delta}$, by ($P3$); it vanishes in equilibrium. Panel (a) is numerically simulated for parameters $\beta_i = \kappa_i = 0.1$ and a logistic distribution $f(x) = e^x/(1+e^x)^2$

there is a conviction if both propose it.[24] Since both jurors have the power to acquit unilaterally while conviction requires consensus, this roughly captures a standard presumption of innocence.

By Lemmas 1–2, jurors follow sincere, agreeable cutoff rules, depicted in Figure 3a. Lones first proposes his natural verdict (convict) if his type ℓ exceeds a threshold x_0, and Moritz opts for his natural verdict to acquit if his type m exceeds some threshold x_1. A lower threshold corresponds to a tougher Lones (or Moritz)—as more types propose their natural verdict.

4.1. *Reaction curves*

The equilibrium cutoffs x_0 and x_1 obey the indifference conditions $\bar{\Pi}_L(x_{-1}, x_0, x_1) = 0$ and $\Pi_M(x_0, x_1, x_2) = 0$. Here, $x_{-1} = -\infty$, for all Moritz' types second Lones if he proposes acquittal, and $x_2 = \infty$ since all Lones' types concede to Moritz in period $t = 2$. Moritz' reaction curve $\Pi_M(x_0, x_1, \infty) = 0$ implicitly yields $m = x_1$ monotonically increasing in x_0. For if Lones hardens his stance by reducing x_0, his conviction proposals are weaker guilt signals. Moritz responds with a tougher stance, *i.e.* a lower acquittal threshold x_1. So Moritz' reaction curve has strategic complements.

In contrast, Lones' (inverse) reaction curve $\bar{\Pi}_L(-\infty, x_0, x_1) = 0$ is "U-shaped" in Figure 3b. To see this, assume first that Moritz' cutoff type x_1 is low. So Moritz acts tough, usually insisting on acquittal, except for very low types $m < x_1$. Here, a conviction proposal by Lones usually delays the verdict. In this case, if Moritz further hardens, by reducing x_1, Lones grows less willing to propose conviction; he softens, raising his cutoff type x_0. Here, Lones' reaction curve exhibits strategic substitutes (lower branch of $\bar{\Pi}_L = 0$ in Figure 3b). At the extreme, when Moritz nearly always asks to acquit (very low x_1), Lones invariably succumbs to Moritz' doggedness, even if Lones is convinced of guilt, since a conviction proposal almost never has any impact.

Assume next that Moritz' cutoff type x_1 is high, corresponding to a soft stance. Then the prospect of a miscarriage of justice is large, and the error of impunity is less likely. Then a

24. This example is a dynamic version of the voting game in LRS. The only difference is that here jurors bear additional waiting cost if Lones proposes convict, but then Moritz insists on acquittal.

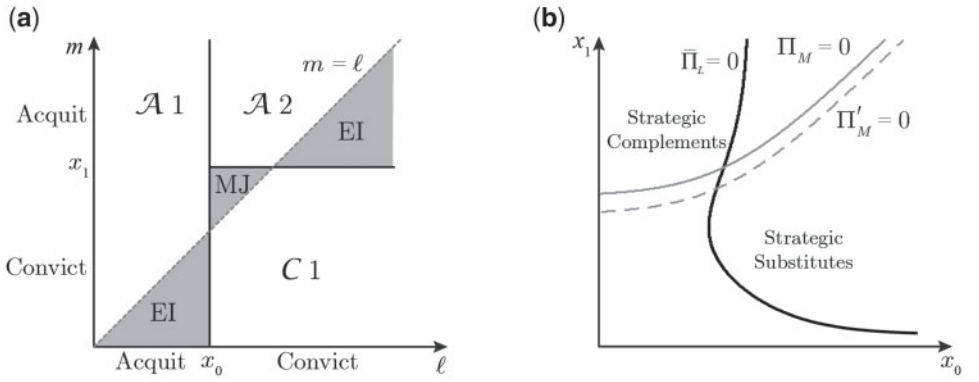

FIGURE 3

Signals: (a) Jury verdicts and (b) reaction curves. Panel (a) depicts jurors' behaviour, debate outcomes and the decision errors given cutoffs x_0, x_1. Acquittal is strictly optimal for $m > \ell$ and conviction for $\ell > m$. We shade the type triangles yielding errors of impunity (EI) and miscarriages of justice (MJ). Panel (b) plots jurors' reaction curves that fix cutoffs x_0, x_1. As Moritz grows more dovish or patient, his reaction curve shifts down from $\Pi_M = 0$ to $\Pi'_M = 0$, as he grows more willing to hold out. Panel (b) is numerically simulated for the signal density and parameters in Figure 2

tougher stance by Moritz (lower x_1) reduces Lones' costs of miscarriages of justice, invoking a more hawkish reply (lower x_0). So Lones' reaction curve exhibits strategic complements.

Figure 3b plots the reaction curves $\Pi_M = 0$ and $\bar{\Pi}_L = 0$. Moritz' reaction curve rises with slope less than one,[25] by Properties $(P1)$ and $(P2)$. Meanwhile, Lones' (inverse) reaction curve, with x_0 as a function of x_1, is "U-shaped" with slope less than one in the upward-sloping branch, by Properties $(\bar{P}2)$ and $(\bar{P}3)$ in Section A.4. Hence, the resulting equilibrium is unique.

4.2. *Comparative statics*

As he grows more biased or patient, Moritz' propensity to hold out for acquittal rises, and so he adopts a tougher stance.

Lones' response depends on whether he exhibits strategic substitutes or complements. If Moritz is initially patient or biased, he acts tough, and usually proposes acquittal. If he grows even more dovish, Lones reacts by softening; for a conviction proposal now simply delays the inevitable acquittal. To wit, Lones' equilibrium cutoff shifts right. His submissive reaction mode is the hallmark of strategic substitutes. In contrast, if Moritz is initially quite impatient or unbiased, he pushes only weakly for acquittal (a high cutoff x_1). In this case, if Moritz grows more patient or biased and proposes acquittal more often, Lones can worry less about miscarriages of justice. He then optimally pushes more strongly for conviction. That greater toughness by Moritz begets a tougher reply by Lones is the signature of strategic complements—as depicted in the upper branch in Figure 3b.

Next, assume Lones grows more biased or patient. He toughens his stance, proposing to convict more. Seeing a weaker signal in the conviction proposal, Moritz pushes to acquit more. That a tougher Lones begets a tougher reply by Moritz reflects his global strategic complements.

25. For $\partial \Pi_M / \partial x_0 < 0$ by $(P1)$ and $\partial \Pi_M / \partial x_0 + \partial \Pi_M / \partial x_1 > 0$ by $(P2)$, when $\Pi_M = 0$. So $\partial \Pi_M / \partial x_1 > -\partial \Pi_M / \partial x_0$. By the Implicit Function Theorem, Moritz' indifference curve has slope $x_1'(x_0) = -(\partial \Pi_M / \partial x_0)/(\partial \Pi_M / \partial x_1) \in (0, 1)$.

4.3. *How delay changes*

Agreement is delayed when Moritz counters a conviction proposal by Lones with an insistence on acquittal. The chance of this event is the probability mass of types northeast of (x_0, x_1) in Figure 3a. If Lones exhibits strategic complements, greater bias or patience of either juror leads each to harden his stance. So cutoffs x_0, x_1 both fall, and delay unambiguously increases. Next assume that Lones exhibits strategic substitutes. If Moritz grows more biased or patient and thereby hardens his stance, Lones softens. All told, x_0 rises, and x_1 falls. With enough bias, Lones' reaction curve grows infinitely elastic (lower portion of Figure 3b), and x_0 rises so much that delay falls. In our focal longer, open-ended equilibria, Propositions 4 and 6 find that *delay rises as either juror grows more biased or more patient*.

4.4. *How decision costs change*

With our presumption of innocence in this truncated equilibrium, errors of impunity are unbounded—the ex post log-likelihood ratio of guilt given an acquittal is unbounded. But consider the maximal log-likelihood ratio of innocence conditional on a convict verdict, $\delta_1 \equiv x_1 - x_0$; in Figure 3b this corresponds to the (horizontal) distance of the intersection of jurors' reaction curves from the 45 degree diagonal (not pictured). Assume that Moritz grows either more dovish or patient. The miscarriage of justice measure falls, as he is more willing to fight such mistakes. If Lones exhibits strategic complements, then he toughens in response. This blunts but cannot reverse the effect of Moritz' tougher stance. But if Lones exhibits strategic substitutes, then he instead softens, further reducing miscarriages of justice.

Next consider either a more hawkish or patient Lones. He pushes harder for conviction, increasing miscarriages of justice. Since Moritz' reaction curve has strategic complements, he toughens in response; this blunts but does not reverse the increase in miscarriages of justice.

In our longer, open-ended equilibria, a more balanced story emerges, since debate need not end in period $\tau = 2$. As Proposition 5 shows later on, if a juror grows more biased or patient, and neither is too biased, then his peer pushes back enough that *losses from both errors fall*.

4.5. *Devil's advocate*

Finally, to flesh out the nature of dynamic debate, we contrast Lones' equilibrium behaviour and his choice were he to call the verdict unilaterally—specifically in the $\sigma = \tau = 1$ deferential equilibrium. Since proposing conviction risks delay, but not with a unilateral decision, one might think that Lones is less inclined to propose conviction in equilibrium. But for small costs the opposite occurs. Namely, Lones becomes a *devil's advocate*, arguing for conviction—despite not wishing that his vote be the last word. For in the deferential equilibrium with $\sigma = 1$ and $\tau = 2$, if Lones proposes acquittal, he seals the verdict, whereas a conviction proposal retains the option to concede when Moritz pushes for acquittal. As long as this option value exceeds the delay costs, Lones pushes against his immediate best interests. Playing the devil's advocate reflects the option value arising in any multi-stage debate setting.

5. AMBIVALENT AND INTRANSIGENT DEBATE

5.1. *Two genres of debate*

Equilibrium balances the costs and benefits of further debate. These include decision gains from winning the debate when winning is ex post optimal, decision costs from winning the debate when losing is ex post optimal, and explicit delay costs. A juror, say Moritz, is *intransigent in period*

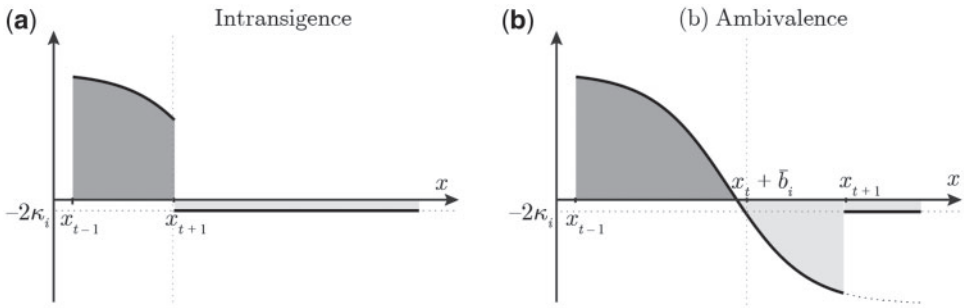

FIGURE 4

Intransigent and ambivalent debate propensities. With intransigence, in panel (a) above, juror i's net decision payoff is positive for conceding peer types $[x_{t-1}, x_{t+1}]$, before jumping down to $-2\kappa_i$. With ambivalence, in panel (b) (and Figure 2a), the net decision payoff is negative for peer types in the interval $[x_t + \bar{b}_i, x_{t+1}]$, before a jump to $-2\kappa_i$

t of the natural debate if—given Lones' strongest type who concedes in the next period x_{t+1}—Moritz' weakest remaining type x_t prefers an immediate acquittal over a conviction one period later: $\Delta(x_{t+1} - x_t, \beta_i) + \kappa_i > 0$. This mimics the incentives in a standard, private value war of attrition, in which players wish to prevail against any opponent's type.

We turn to our more novel genre of debate. For debate need not be win-lose. Call Moritz *ambivalent in period t* if—given Lones' strongest conceding type next period—he prefers a conviction next period over an immediate acquittal, *i.e.* if the expected decision payoff of his natural verdict plus one period's waiting cost is negative: $\Delta(x_{t+1} - x_t, \beta_i) + \kappa_i < 0$. When Moritz is ambivalent, he is of two minds, keenly aware that his vote may be a mistake. In this taxonomy, Moritz is naturally intransigent or ambivalent in a period.

By (10) and our definitions of intransigence and ambivalence, a juror is

$$\text{juror } i(t) \text{ is } \begin{cases} \textit{intransigent in period } t \text{ if } x_{t+1} < x_t + \bar{b}_i, \\ \textit{ambivalent in period } t \text{ if } x_{t+1} \geq x_t + \bar{b}_i. \end{cases} \tag{11}$$

Put differently, a juror, say Moritz, is intransigent in period t when few strong types of his peer Lones concede in period $t+1$, for then $x_{t+1} - x_t < \bar{b}_M$. But as the marginal conceding type x_{t+1} of Lones grows, Moritz transitions into ambivalence, where $x_{t+1} - x_t > \bar{b}_M$.

We now reformulate our debating genres in terms of the jurors' best reply functions. If, say, Moritz is intransigent in some period, then a tougher stance by Lones in the next period (so fewer types conceding) reduces Moritz' propensity to hold out. For it reduces his decision payoff gain from winning the debate against weak types of Lones, by shrinking the positive part of the integral in Figure 4a. This begets a weaker reply by Moritz (greater x_t)—to wit, local strategic substitutes. Conversely, if Moritz is ambivalent in period t, a tougher stance by Lones in the next period raises Moritz' propensity to hold out, by cutting his decision payoff losses from winning the debate against strong types of Lones—the negative integral portion in Figure 4b. This begets a tougher reply by Moritz in period t; that is, local strategic complements.

Figure 5 depicts the two debate genres.[26] The *disagreement zone* is all pairs (ℓ, m) with $\ell \leq m + \bar{b}_M$ and $m \leq \ell + \bar{b}_L$, *i.e.* where jurors disagree about the best verdict, up to one period's

26. We depict the genres separately. But in fact, Proposition 3 will argue that the genre may transition over time, and asymptotically the two jurors can be engaged in different debate genres (see Section 7).

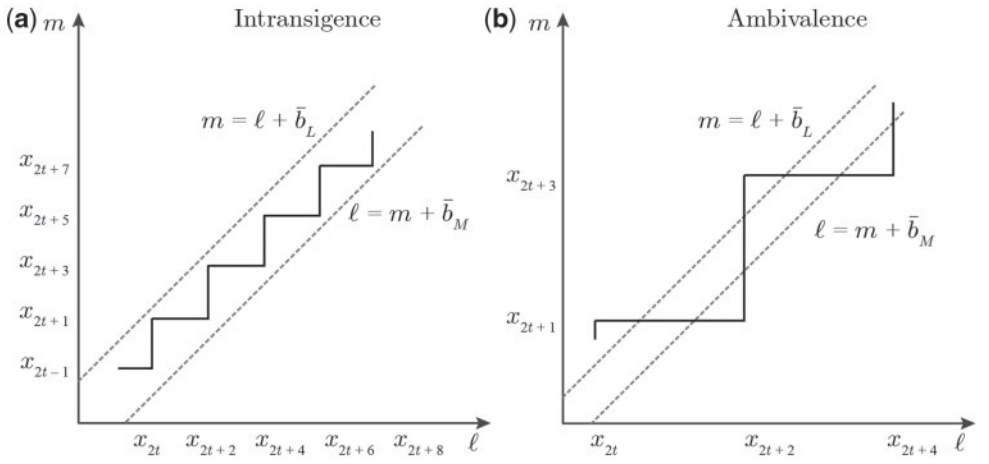

FIGURE 5

Intransigent and ambivalent debate: cutoff vectors. Jurors disagree in the disagreement zone between the two indifference lines. These lines collapse onto the diagonal as biases and waiting costs vanish (by (10)). Since only weak types of the rival juror concede in intransigent debate, jurors disagree after the debate; graphically, cutoffs are bracketed by these lines in this case. And because strong types of the rival juror concede in ambivalent debate, jurors agree at the end of debate; in this case, cutoffs straddles two indifference lines

waiting costs. For intransigent debate in Figure 5a, consecutive cutoff pairs lie inside the disagreement zone. The debate speed, as measured by the staircase step size, is bounded above by the width of the disagreement zone. In contrast, for ambivalent debate in Figure 5b, cutoff pairs zick-zack around the disagreement zone; this bounds the speed of debate from below.

While intransigent debate outcomes are win–lose, ambivalent debate can yield inefficient verdicts—*i.e.*, convictions in the triangles above the disagreement zone where even Lones prefers acquittals, and acquittals in the triangles below, where even Moritz prefers conviction. We measure these decision errors from an unbiased observer's perspective. Call the *cutoff gap* $\delta_{2t+1} \equiv x_{2t+1} - x_{2t}$—the log-likelihood ratio of innocence if Moritz' cutoff type x_{2t+1} concedes to Lones' prior cutoff type x_{2t}—the (maximal) *miscarriage of justice measure* in period $2t+1$, and the cutoff gap $\delta_{2t} \equiv x_{2t} - x_{2t-1}$ is the (maximal) *error of impunity measure* in period $2t$.

5.2. *Incidence of debating genres*

We now explore how the debate genre depends on jurors' bias, delay costs, and information.

The debate genre may differ across jurors and vary over the course of the debate. But we will see that for extreme bias or patience, the genre is unchanging throughout the debate. We say that *natural debate is ambivalent* (resp. *intransigent*) if both jurors are ambivalent (resp. intransigent) in all periods of the natural debate; we similarly describe Nixon-China debate.

For starters, let Moritz be a partisan who prefers to acquit all defendants, even those whom he knows are guilty: $\beta_M \geq 1$ (ruled out). Then, much like in a standard, private value war of attrition, he never wishes that Lones overturn his proposal and natural debate is intransigent.

We argue that debate is ambivalent at the opposite extreme, with perfectly aligned interests: no bias $\beta_L = \beta_M = 0$ and identical delay costs $\kappa_L = \kappa_M$. Indeed, for a contradiction, suppose that Moritz' cutoff type x_t intransigently wishes to prevail and acquit. With zero delay costs, Lones' cutoff type x_{t+1} conditions on even stronger evidence for innocence and strictly prefers to concede

in period $t+1$; for Lones only know that Moritz' type *exceeds* x_t, since Moritz still remains in the debate. This contradicts the indifference of Lones' type x_{t+1}—whence Moritz must have been ambivalent in period t. This logic persists for all common delay costs $\kappa_L=\kappa_M>0$. For in this case, $\bar{b}_L=\bar{b}_M=-\underline{b}_L=-\underline{b}_M$ from (10). If Moritz were intransigent, then (11) implies $\delta_{t+1}=x_{t+1}-x_t<\bar{b}_M=-\underline{b}_L$. By property $(P1)$, Lones' gap propensity $\pi_L(\delta_{t+1},x_{t+1},\delta_{t+2})<0$ for any δ_{t+2}, and his cutoff type x_{t+1} strictly wants to concede; contradiction.

Lemma 3. (Negative Propensity Proviso). *Natural debate is ambivalent provided:*

$$\pi_M(\bar{b}_L,y,\bar{\delta})<0 \text{ and } \pi_L(\bar{b}_M,y,\bar{\delta})<0 \text{ for all real } y,\bar{\delta}. \tag{12}$$

To understand Lemma 3, assume for a contradiction that Moritz, say, is intransigent in period t; *i.e.* given Lones' strongest type $\ell=x_{t+1}$ who next concedes, Moritz' weakest remaining type x_t prefers an immediate acquittal over a conviction one period later. Shooting ahead one period, we argue that Lones' cutoff type x_{t+1} strictly wants to concede, thereby contradicting Moritz' intransigence in period t. Lones' benefit of holding out consists of the net decision payoff gain from securing a conviction against weak types of Moritz $m<x_{t+1}+\underline{b}_L$—recalling that Lones' net decision payoff tips negative at $x_{t+1}+\underline{b}_L$ in Figure 2. Since Moritz' types $m<x_t$ have already conceded, and $x_t\geq x_{t+1}-\bar{b}_M$ by (11), the interval of Moritz' types for whom Lones' net decision payoff is positive has length at most $\bar{b}_M+\underline{b}_L$. The second inequality in (12) then implies that Lones' propensity is negative, irrespective of Moritz' next cutoff x_{t+2}; this contradiction implies that Moritz must be ambivalent in period t.

The premise of Lemma 3 follows if $\bar{b}_L=\bar{b}_M=-\underline{b}_L=-\underline{b}_M$, by property $(P1)$. By continuity, the premise reassuringly holds for approximately common interests, *i.e.* when biases β_i are small enough and delay costs κ_i are near enough—as we prove in Section A.7.[27] The premise of Lemma 3 holds in particular for the parameters and density described in Figure 2 (numerically verified).

The proviso (12) in Lemma 3 also holds when jurors' signals are informative enough. To see this, fix Moritz' type y. Loosely, the more informed is Lones—*i.e.* the larger is the inverse limit hazard rate— then the larger is his random type ℓ. In other words, the density f increasingly assigns probability weight in (9) to large gaps δ (Figure 2a) with a negative net decision payoff $\Delta(\delta,\beta_i)-\kappa_i<0$ or negative delay costs $-2\kappa_i$—respectively, for $\delta<\bar{\delta}$ or $\delta>\bar{\delta}$. This gives the first inequality of (12); the second owes to a symmetric argument for Lones. Summarizing:

Proposition 1. (Ambivalence). *Natural debate is ambivalent for sufficiently close interests or informative types. With symmetric jurors, natural debate is ambivalent for small bias* $\beta\geq 0$.[28]

The logic adjusts in the Nixon-China subgame, where the hawk Lones argues for acquittal against his natural position, and the dove Moritz pushes for conviction. Here, ambivalence arises far more readily for biased jurors. To see why, let us revisit the ambivalence logic for Proposition 1. Namely, suppose that Moritz intransigently wishes to prevail, *i.e.* to achieve his *unnatural* conviction verdict. In the natural debate, we leveraged the *bias upper bound* to derive the contradiction that Lones' subsequent cutoff type strictly wishes to concede. For Nixon-China

27. Assumption $(\star)$ in Section 3.5 plays a logical role here. For assume instead that f has a vanishing inverse hazard rate, *i.e.* $\gamma=0$. Assume a conflict of interest, precluding $\beta_L=\beta_M=0$ and $\kappa_L=\kappa_M$. Then $-\bar{b}_L<\underline{b}_M$, by (10). In this case, inequalities (12) fail for large y. Consider the first inequality. Indeed, the conditional density $f(y+\delta)/F(y-\bar{b}_L)$ assigns most probability to gaps δ close to $-\bar{b}_L$ where the net decision payoff $\Delta(\delta,\beta_M)-\kappa_M>0$, since it decreases and vanishes at $\underline{b}_M>-\bar{b}_L$. So $\pi_L(\bar{b}_L,y,\bar{\delta})>0$ for any $\bar{\delta}>-\bar{b}_L$ and large y.

28. While Theorems are purely technical results, the Propositions contain our substantive economic predictions.

debate, where any concession by Lones leads to conviction, a *bias lower bound* suffices—since Lones' hawkishness *reinforces* his willingness to concede. We accordingly prove in Section A.8 that:

Proposition 2. *Nixon-China debate is ambivalent if preference bias dominates cost divergence, $|\beta_L+\beta_M|\geq|\kappa_L-\kappa_M|$, and thereby is ambivalent if $\kappa_L=\kappa_M$, and thus for symmetric jurors.*

We return to the natural subgame. Here, bias and delay costs intuitively work at cross-purposes in jurors' preferences: A biased juror more eagerly holds out, while a more impatient juror more eagerly concedes. Since Proposition 1 finds that low bias leads to ambivalence, this suggests that intransigence arises with larger bias, or equivalently, small delay costs. To be more precise, write the per period delay cost $\kappa_i=k_i\eta$, where k_i is juror i's *flow waiting cost* and $\eta>0$ the *real period length*. Now, fix the bias $\beta_i>0$ and flow waiting costs $k_i>0$ for $i=L,M$.

Proposition 3. (Intransigence). *For any real time $\tau>0$ and short enough real time period lengths $\eta>0$, natural communicative debate is intransigent after period $t^*(\eta)\equiv\lfloor\tau/\eta\rfloor$. Also, the per period hazard chance of debate ending after period $t^*(\eta)$ is of order η, as $\eta>0$ vanishes.*

Let us flesh this out. While the ambivalence analysis in Proposition 1 applies to all equilibria, the intransigence result only applies to communicative equilibria. For in any τ-deferential equilibrium in the natural subgame, whoever moves in period $\tau-1$ is ambivalent—as he surely does not wish to prevail against the strongest conceding types of his peer juror (recalling (11)). Also, intransigence in a communicative equilibrium may be preceded by an initial ambivalent debate phase. For we must bridge the ambivalent Nixon-China debate, mandated by Proposition 2. All told, types above a threshold x^* are intransigent, where $x^*\to-\infty$ as the period length $\eta\downarrow 0$.[29]

For the last claim of Proposition 3, assume that the period length $\eta>0$ and the delay costs $\kappa_i=\eta k_i$ both vanish. Then in Figure 4a, the negative integral vanishes (its width is $2\kappa_i$), and therefore the positive part also vanishes, by optimality. So a vanishing type interval $[x_{t-1},x_{t+1}]$ concedes every period,[30] and the hazard rate of debate ending in any given period vanishes, too. This is analogous to a standard war of attrition with vanishing per-period delay costs.

Propositions 1–3 jointly characterize debate by symmetric jurors with delay costs $\kappa>0$ and bias $\beta\geq 0$. In a communicative equilibrium, Nixon-China debate is ambivalent, while natural debate is ambivalent for small bias (given delay costs), but transitions into intransigence for small delay costs (given the bias). While Propositions 1 and 3 may appear to conflict for small delay costs and biases, they do not. For the negative propensity proviso (12) holds—debate is ambivalent—for fixed common delay costs $\kappa_L=\kappa_M>0$ and small biases $\beta_L,\beta_M\geq 0$, but fails for fixed biases $\beta_L,\beta_M>0$ and small delay costs $\kappa_L,\kappa_M>0$ (short period lengths $\eta>0$).[31]

29. Proposition 3 is silent on the initial debate genre. Since the number of periods in real time $\tau>0$ explodes as $\eta>0$ vanishes, one might worry that debate may terminate before hitting the intransigent phase (thinking of the Coase Conjecture). In fact, the proof of Proposition 3 in Section A.9 shows that, as τ and η vanish, the chance that debate lasts at least real time $\tau>0$ tends to one, and so debate almost always is eventually intransigent. Moreover, the expected duration of intransigent debate is boundedly positive, since the hazard rate of debate ending by any period t vanishes in η (Proposition 3). Since the real time before debate turns intransigent τ vanishes in the period length $\eta>0$ (Proposition 3), and the probability of natural debate tends to one (as we show in Section A.9), an outside observer witnessing the debate typically sees intransigent debate.

30. Algebraically, this follows by substituting (4) into (6); as the second integral in (4) vanishes and the integrand of the first integral is bounded away from zero, the domain of the first integral must vanish.

31. Indeed consider, say, Lones' propensity $\pi_L(\bar{b}_M,y,\bar{\delta})$ for any type y, upper gap $\bar{\delta}=\bar{b}_L$ and *zero* delay costs ($\kappa_L=0$). The net decision payoff—the first term of (9) depicted in Figure 2a—is then positive on its entire domain, as it

For an instructive counterpoint to our analysis, consider the static voting game of LRS. Our twin insights for ambivalent equilibrium that (1) even the dove Moritz worries about the error of impunity when proposing to acquit, and (2) rational debate may entail *ex post* inefficient acquittals are reminiscent of LRS' static voting model. Our dynamic model approximates LRS when delay costs vanish.[32] The equilibria solely feature ambivalence, as indifference demands balancing the two kinds of decision errors. The possibility of ex post inefficient verdicts ensures that jurors do not wish to magnify the strength of their signals. But our model with vanishing delay costs instead ensures honest juror voting with delay costs. Thus, small delay costs is quite unlike the zero delay cost limit, because delay costs can be endogenously amplified in equilibrium. Formally, the equilibrium correspondence fails upper hemi-continuity at $\eta=0$.[33] We argue in Online Appendix Section B.2 that the limit of our discrete time model with vanishing delay costs is a continuous time model, which features intransigence in the communicative equilibrium.

6. AMBIVALENT DEBATE

We now explore ambivalent natural debate, which arises for sufficiently unbiased or informed jurors (Proposition 1), since it is our core novel contribution. We first argue that equilibria are unique under the negative propensity proviso (12), and then derive comparative statics.

6.1. *The unique equilibrium*

We argue that the second-order difference equation defined by (6) and (8) admits a unique solution. We present our argument for natural debate here, since it is technically innovative, and takes inspiration from saddle point proofs in optimal control.

We rewrite the indifference condition (6) for natural debate in terms of equilibrium cutoff gaps, namely $\pi_i(\delta_t,x_t,\delta_{t+1})=0$, and solve it forwardly for δ_{t+1}. As in Section 5, by property $(P3)$ in Section 3.5, ambivalence requires $\delta_{t+1}>\bar{b}_i$. The indifference condition admits at most one such root,[34] denoted $\chi_i(\delta_t,x_t)$ when it exists. Properties $(P1)$–$(P3)$ then imply that the *shooting function* χ_i increases in its arguments; more strongly, the slope of χ_i in δ_t exceeds $1+\varepsilon$ for large x_t, by $(P5)$.

For any "anchor" x_0, the "seed" x_1 then recursively fixes the cutoff sequence $x_2,x_3,\ldots$ by iterating $\delta_{t+1}=\chi_{i(t)}(\delta_t,x_t)$. The τ-deferential equilibrium has the boundary condition $x_\tau=\infty$, whereas the communicative equilibrium obeys the transversality condition $x_t\to\infty$. We claim that this cutoff sequence increases in its seed. For if $x_1'>x_1$, then $\delta_1'=x_1'-x_0>x_1-x_0=\delta_1$, and thus

$$\delta_2'-\delta_2=\chi_M(\delta_1',x_1')-\chi_M(\delta_1,x_1)>0 \tag{13}$$

hence $x_2'=x_1'+\delta_2'>x_1+\delta_2=x_2$; inductively, $x_t'>x_t$ for all t. So, for *finite* τ, there is at most one seed x_1 with $x_\tau=\infty$. Uniqueness of the sequence follows whenever $\tau<\infty$.

is decreasing and vanishes at the upper end of the interval $\bar{\delta}=\underline{b}_L=\bar{b}_L$, where $\underline{b}_L=\bar{b}_L$ because $\kappa_L=0$ in (10). The second delay cost term of (9) vanishes, too, and so the gap propensity $\pi_L(\bar{b}_M,y,\bar{b}_L)>0$, that is (12), fails. This logic extends by continuity to small $\kappa_L=k_L\eta>0$.

32. More precisely, the normal-form of our model and LRS' multi-vote extension in their §V share a common limit as our period length $\eta\downarrow 0$, and their maximum number of votes explodes. In this limit game, Lones chooses an odd period $t\in 2\mathbb{Z}+1\cup\{\pm\infty\}$, Moritz chooses an even period $s\in 2\mathbb{Z}\cup\{\pm\infty\}$, the defendant is convicted if $t>s$, acquitted if $s<t$, and payoffs are $-\infty$ if $s=t=\infty$ or $s=t=-\infty$.

33. No sequence of intransigent equilibria for vanishing η converges to an equilibrium in this limit. This failure of upper hemi-continuity owes to our violation of continuity at infinity (Fudenberg and Tirole, 1991). Conversely, deferential equilibria approximate every equilibrium in LRS' multi-vote model as $\eta\downarrow 0$.

34. Without the ambivalence inequality (12), there could be multiple roots of the indifference condition $\pi_i(\delta_t,x_t,\cdot)=0$ and it is no longer clear how to argue equilibrium uniqueness.

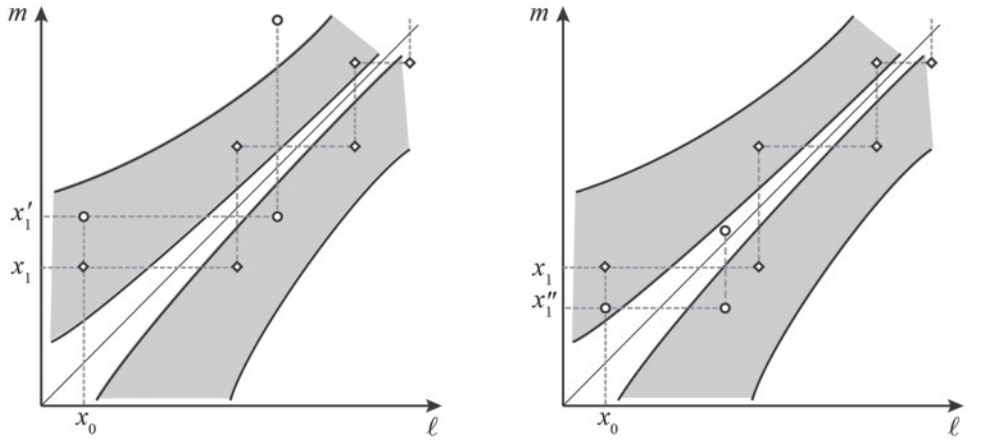

FIGURE 6

Equilibrium cutoffs as a dynamical system. The anchor x_0, seed x_1, and shooting function $\chi_{i(t)}(x_{t-1},x_t)=x_{t+1}$ jointly define a sequence of possible debate conclusions $(x_1,x_2),(x_3,x_2),\cdots$ (diamonds). The sequence of circles, defined by the alternative seed $x_1'>x_1$ in (a), "fans out" and eventually leaves the domain of $\chi_{i(t)}$ on the outside. The sequence of circles defined by alternative seed $x_1''<x_1$ in (b), "fans in" and leaves the domain on the inside

Uniqueness of the communicative equilibrium is more subtle since two cutoff sequences (x_t) and (x_t') may conceivably explode at different rates. So inspired, strengthen (13) to $\delta'_{t+1}-\delta_{t+1}>(1+\varepsilon)(\delta'_t-\delta_t)$ for large t, since the slope of χ_i in δ_t eventually exceeds $1+\varepsilon$. For intuitively, if juror $i(t)$ convinces more weak peer types in (x_t') than in (x_t) (a greater, positive *DG* area in Figure 2a), then extra decision payoff gains must be balanced by extra decision payoff costs (a greater, negative *DL* area in Figure 2a). As there are more weak than strong peer types with the falling density, the cutoff gap difference increases. But then $(\delta'_{t+s}-\delta_{t+s})>(1+\varepsilon)^s(\delta'_t-\delta_t)$ diverges, and the cutoff gaps δ_t' perforce explode too. This "fanning out", seen in Figure 6a, is inconsistent with a communicative equilibrium, since the positive area in Figure 2a swamps the negative area for large enough lower gaps $\underline{\delta}$, given the exponentially vanishing density $\underline{f}$. Altogether, for large enough cutoff gap δ_t', the gap propensity $\pi_i(\delta_t',x_t',\delta_{t+1})>0$ for any $\delta_{t+1}>\bar{b}_i$.

In summary, for any initial cutoff x_0, and any finite or infinite drop-dead date τ, the second order difference equation for $(x_t)_{t\geq 0}$ has a unique solution. Finally, in Section A.10, we solve for the unique equilibrium value x_0 by application of a similar logic to the Nixon-China subgame, and a monotonicity argument for the initial propensity (7). We conclude:

Theorem 5. *Assume the ambivalence inequalities (12). Then each (σ,τ)-equilibrium is unique.*

The earlier fanning out argument also implies that both cutoff gaps δ_{2t} and δ_{2t+1} strictly fall over time as communicative debate transpires, and thus converge:[35] To see this, assume for a contradiction that $\delta_{t+2}\geq\delta_t$ for some t. Consider the shifted cutoffs $x_t'\equiv x_{t+2}$ and cutoff gaps $\delta_t'=\delta_{t+2}$, so that $\delta_t'\geq\delta_t$. Since some types concede in every period of a communicative debate, the cutoffs strictly increase, whence $x_t'>x_t$. Then the second difference $\delta_{t+2}-\delta_t\equiv\delta_t'-\delta_t$ grows exponentially; this is inconsistent with equilibrium, as argued before Theorem 5. Given our focus

35. But decreasing gaps need *not* imply that debate slows down. For the distribution f is log-concave, and so constant cutoff gaps mean that debate speeds up.

on small biases and thus ambivalent debate, (11) implies that the (decreasing) decision errors are also bounded below by $\bar{b}_L$ and $\bar{b}_M$, and so converge. As we assume an asymptotically stationary type distribution f, the hazard rate of communicative natural debate also settles down.

6.2. *Equilibrium predictions*

We now explore how the unique (σ,τ)-equilibrium behaviour reflects fundamentals. We say that natural debate *slows down* when the cutoffs uniformly fall, from (x_t) to (x'_t), where $x'_t < x_t$ for all $t > 0$. For then natural debate is less likely to end before period t. Likewise, the Nixon-China debate slows down if $x'_t > x_t$ for all $t < 0$.

Proposition 4. (Greater Bias or Patience). *Assume the negative propensity proviso (12). If juror i grows more biased (β_i rises), then in any (σ,τ)-equilibrium, natural debate slows down, while Nixon-China debate speeds up. With unbiased jurors, $\beta_i = 0$, if either juror grows more patient (κ_i falls), then both natural and Nixon-China debate slow down in any (τ,τ)-equilibrium.*

We see here the impact of increased polarization in debate: When jurors grow more biased and steadfast in their positions, natural debate slows down; and they concede more slowly. The impact of prior bias on debate accuracy is important for the analysis of juries and panels. Because the juror discussion grows more fine-grained, the decision errors intuitively fall. While a proof is only possible for the asymptotic analysis in Section 7, this stronger conclusion is consistent with the example in Section 4.[36] Polarization also impacts the chance of natural and Nixon-China debate. For when $\beta_L = \beta_M = 0$, symmetry demands $x_0 = 0$. We show in Section A.10 that the initial cutoff x_0 falls in β_L or β_M. To wit, it is more likely that debate is natural with greater polarization.

Note that we can speak to the impact of patience in the case of two unbiased jurors. Here, when Lones' delay costs drop, he is more willing to delay, and thus debate ends later.[37]

7. ASYMPTOTIC DEBATE

We now secure our sharpest predictions for asymptotic communicative debate. We argue that communicative debate eventually settles down, owing to the asymptotically exponential type distribution. So motivated, call the debate *asymptotically stationary* if the sequences of cutoff gaps $(\delta_{2t}),(\delta_{2t+1}),(\delta_{-2t})$, and (δ_{-2t-1}) converge.[38] When their limits exist, call them $\delta_{MJ} \equiv \lim_{t\to\infty}\delta_{2t+1}$ and $\delta_{EI} \equiv \lim_{t\to\infty}\delta_{2t}$, and for the Nixon-China subgame, $\hat{\delta}_{MJ} \equiv \lim_{t\to-\infty}\delta_{-2t+1}$ and $\hat{\delta}_{EI} \equiv \lim_{t\to-\infty}\delta_{-2t}$. Given the interpretation of cutoff gaps as decision errors, call δ_{MJ} the *eventual miscarriage of justice*, and δ_{EI} the *eventual error of impunity*.[39]

Theorem 6. *Communicative debate is uniquely asymptotically stationary: Limit cutoff gaps δ_{MJ}, δ_{EI}, $\hat{\delta}_{MJ}$, and $\hat{\delta}_{EI}$ are well-defined and unique.*

36. Complementing Proposition 4, Online Appendix Section B.4 discusses how equilibrium cutoffs x_t, and hence delay, vary in the informativeness of jurors' signals.

37. Our proof shows that *for a fixed initial cutoff* x_0, the cutoff vector $x_1,\ldots,x_\tau$ consistent with a τ-equilibrium in the natural debate increase in κ_i (and similarly for Nixon-China debate). When jurors are unbiased, symmetry and equilibrium uniqueness anchors the initial cutoff at $x_0 = 0$, and so debate speeds up. When jurors are biased, our analysis is marred by an indeterminate shift in x_0 as κ_i increases.

38. Indeed, we saw after Theorem 4 for small biases, that alternating cutoff gaps converge monotonically.

39. Since the sum of cutoff gaps $\delta_{t+1} + \delta_t = x_{t+2} - x_t$ is positive, so too is its limit $\delta_{MJ} + \delta_{EI}$. But either gap by itself and its limit, δ_{MJ} and δ_{EI}, may be negative.

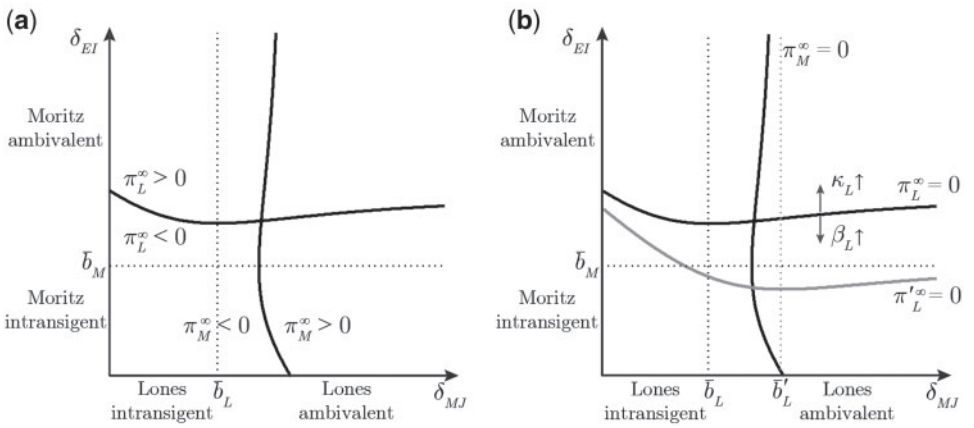

FIGURE 7

Limit gaps. Panel (a) depicts the asymptotic indifference curves, separating the regions where jurors prefer to hold out and concede. The shape of the curves reflects properties $(P1)$, $(P3)$ and $(P5)$. Panel (b) depicts how Lones' curve shifts down and the limit gaps shift down on Moritz' inversely U-shaped curve, as β_L rises or κ_L falls. The simulation in this figure assumes $\beta_i = \kappa_i = 0.1$, $\beta_L = 0.3$ and the logistic type density $f(x) = e^x/(1+e^x)^2$

For recall the limit propensity π_i^∞ from Section 3.5. The equilibrium limit gaps obey the *asymptotic indifference curves* $\pi_M^\infty(\delta_{MJ}, \delta_{EI}) = 0$ and $\pi_L^\infty(\delta_{EI}, \delta_{MJ}) = 0$, when the first argument is an implicit function of the second. By properties (P1) and (P3), and then (P5), we next deduce:

$(I1)$ The asymptotic indifference curves are U-shaped and have less than unit absolute slope.

By this property, the asymptotic indifference curves intersect at most once, as in Figure 7a.

We now pursue equilibrium predictions for the unique limit cutoff gaps. Recall from Section 5 that Lones is intransigent in period $2t$ for small cutoff gaps $\delta_{2t+1} = x_{2t+1} - x_{2t} < \bar{b}_L$, and ambivalent for large cutoff gaps $\delta_{2t+1} > \bar{b}_L$. Then in communicative natural debate, Lones is *eventually intransigent* if $\delta_{MJ} < \bar{b}_L$ and *eventually ambivalent* if $\delta_{MJ} > \bar{b}_L$. Moritz is likewise eventually intransigent or ambivalent as $\delta_{EI} \lessgtr \bar{b}_M$. Our next result exploits these thresholds.

Proposition 5. (Decision Errors). *As Lones grows more biased or patient, the eventual error of impunity δ_{EI} decreases; the eventual miscarriage of justice δ_{MJ} falls if Moritz is eventually ambivalent and rises if he is eventually intransigent. Symmetrically, as Moritz grows more biased or patient, the eventual miscarriage of justice δ_{MJ} decreases, and the eventual error of impunity δ_{EI} falls if Lones is eventually ambivalent, and rises if he is eventually intransigent.*

Figure 7b gives a graphical proof. For the indifference curve $\pi_i^\infty = 0$ satisfies the opposite monotonicity of π_i^∞; that is, by property $(P4)$, it falls in β_i and increases in κ_i, as in Figure 7b.

$(I2)$ The asymptotic indifference curves falls in the bias β_i and increases in the waiting cost κ_i.

A more hawkish or patient Lones pushes harder for convictions, and naturally reduces the eventual error of impunity δ_{EI}. But surprisingly, if Moritz is eventually ambivalent ($\delta_{EI} > \bar{b}_M$), then the corresponding eventual miscarriage of injustice δ_{MJ} falls as well—for then Lones'

tougher stance leads Moritz to push harder for acquittals.[40] This contrasts with the cheap talk literature (Crawford and Sobel (1982) and LRS), where greater bias leads to *greater* decision errors. As Lones grows more biased or patient, we may reach a tipping point, as Moritz shifts into eventual intransigence.[41] At this point, Lones concedes so few eventual errors of impunity that Moritz wishes to prevail against all conceding types of Lones. A more patient or biased Lones then erodes Moritz' propensity to push for acquittals, and raises the eventual miscarriage of justice δ_{MJ}. So as one juror grows more biased or patient, *his peer* single-crosses from eventual ambivalence into intransigence, as seen in Figure 7b.[42] This corroborates our insights in Section 5 that strongly biased jurors are intransigent, while unbiased jurors are ambivalent in all periods.[43]

We can now address a concrete question: What happens if fewer peremptory challenges are afforded lawyers during the jurors' voire dire? This limits their ability to strike opinionated jurors. To capture this change, we posit more biased jurors. Assume symmetric jurors, with a common bias $\beta \geq 0$ and delay costs $\kappa > 0$, so that $\bar{b}_L = \bar{b}_M = \bar{b}$ in (10). As bias increases, the jurors' asymptotic indifference curves in Figure 7b shift down and left, respectively, and their crossing point—the eventual decision errors—shifts down on the 45-degree line: To wit, more symmetrically biased jurors make better decisions.[44] Crucially, the opposite message emerges in the static cheap talk literature. But in our dynamic setting, *symmetric bias increases information revelation, and so leads to better decisions*. Since peremptory challenges reduce bias, they lead to worse jury decisions. This offers a strategic rationale for England's abolition of these challenges in their Criminal Justice Act of 1988.

While the two eventual decision errors can move in opposite directions, one can obtain sharper predictions by considering their sum $\delta_{MJ} + \delta_{EI}$, which measures the total decision error. As Lones, say, grows more biased or patient, his asymptotic indifference curve in Figure 7(b) drops by property ($I2$); therefore, the sum $\delta_{MJ} + \delta_{EI}$ falls since Moritz' (inverse) asymptotic indifference curve has absolute slope less than one, by property ($I1$). Intuitively, the direct effect of a more hawkish Lones pushing harder is only partially countered by the indirect effect of Moritz pushing less hard to acquit. This result explains how the biased advocacy foundation of the adversarial trial system can secure smarter verdicts. For instance, appeals courts in the U.S.—whose judges are alternately appointed by Democratic and Republican administrations—may reach smarter decisions precisely due to their partisan nature.

We next consider the asymptotic speed of debate. Since the type density is asymptotically exponential and cutoff gaps converge, the *eventual hazard rate of ending natural debate*[45]—an increasing function of $\delta_{MJ} + \delta_{EI}$ of the conceding type interval—intuitively also converges.

40. In the two-period example in Section 4, Moritz pushes harder for acquittals as Lones grows more biased, and yet miscarriages of justice *increase*. This difference reflects two ways that a more biased Lones raises Moritz' asymptotic propensity: He prevails over more weak types *and* fewer strong types of Lones. The second effect is absent in the two-period example where all Lones' remaining types concede in period two.

41. Figure 7b shows the case where the tipping point is reached. But since the limit gaps are a function of the parameters $\kappa_L, \kappa_M, \beta_L, \beta_M$, there may be parameter values β_M, κ_M for Moritz, such that when we vary Lones' parameters κ_L or β_L, the equilibrium limit gaps trace only the ambivalent branch of Moritz' curve (or only its intransigent branch). For example, unbiased debate is ambivalent for any delay costs, while highly biased debate is intransigent for any delay cost, as noted in Section 5.

42. In fact, Figure 7b strongly suggests that Lones also switches from ambivalence to intransigence as he grows more biased since the threshold $\bar{b}_L$ shifts right; while this is intuitive, we cannot prove this property in general.

43. Complementing Proposition 5, Online Appendix Section B.4 discusses how limit gaps δ_{MJ}, δ_{EI}, and eventual decision errors, vary in the informativeness of jurors' signals.

44. The shift in eventual decision errors was ambiguous in Proposition 5 where only one juror's preferences shift.

45. Namely, the chance of debate ending in period $2t$ (for Lones) or $2t+1$ (for Moritz) conditional on reaching these late periods; formally, these equal $\lim_{t\to\infty} F(x_{2t} | y \geq x_{2t-1}, x \geq x_{2t-2})$ and $\lim_{t\to\infty} F(x_{2t+1} | y \geq x_{2t}, x \geq x_{2t-1})$.

Proposition 6. (Debate Speed). *When either of the jurors grows more biased or more patient, the eventual hazard rate of ending the debate falls.*

Whereas Proposition 4 found that the *length* of debate rises in bias and patience, here we assert that the *hazard rate* of debate falls late in the communicative equilibrium. This complements our finding in Section 5 that the hazard rate of debate vanishes in the period length—and with it, the per period delay cost vanishes. Proposition 6 follows from our above observation that the sum of decision errors $\delta_{MJ}+\delta_{EI}$ falls when either juror grows more patient or biased. In the two-period example in Section 4, we saw how debate may speed up when a juror grows more biased. In contrast, here we can conclude that debate slows down. Usefully, this allows us to identify the debate genre from its length: For greater bias or patience leads to debate that is intransigent (by the discussion following Proposition 5) and long (by Proposition 6). Thus, long debates indicate intransigence and, conversely, short debates indicate ambivalence.

A distinctly different lesson emerges for Nixon-China debate:

Proposition 7. (Nixon-China). *As either juror grows more biased or less patient, the eventual decision errors $\hat{\delta}_{MJ}, \hat{\delta}_{EI}$ and the eventual hazard rate of debate ending increase.*

To understand this result, imagine the (missing) Nixon-China analogue of Figure 7b. The only qualitative difference is that the Nixon-China asymptotic indifference curve $\hat{\pi}_i^{\infty}=0$ *shifts up* when β_i rises; formally this follows from property ($\hat{P}4$) of the Nixon-China limit propensity $\hat{\pi}_i^{\infty}$ in Section A.4. Intuitively, a more biased juror is less willing to argue for his unnatural verdict, and so his indifference requires convincing a larger interval of his opponent's types. By strategic complementarity, as either juror grows more biased, the eventual decision errors in the Nixon-China subgame both increase, and so too does the eventual hazard rate of debate ending.

In sum, Propositions 6–7 deliver opposite predictions for the two subgames as jurors grow more biased, but both agree that debate slows down in both subgames as jurors grow more patient. This contrast allows one in principle to identify the parameters β and κ in our model.

8. OPTION VALUE OF DEBATE

Unlike a standard, private-value war of attrition where each player always wishes to prevail, our jurors may be convinced by a peer's repeated arguments. We now argue that even once a juror's private posterior tips in favour of the opposite verdict, he may yet persist in his disagreement.[46] To see this, consider the *dictator's problem*: how should one non-omniscient juror vote if he were suddenly given dictatorial power and asked to decide unilaterally. This differs from our *debater's problem*, where the refusal to concede is never final, and each period of delay costs a juror. Since the debater must pay a deliberation cost but the dictator need not, one might think that the debater is more willing to concede than the dictator. But for small biases and waiting costs, just the opposite occurs: Our jurors become *devil's advocates*, arguing for their less preferred verdict. The refusal to concede is prized for its option value, namely, for the additional information it reveals. For seconding a proposal ends the game, but holding out retains the option value of conceding later based on additional information about the peer's type. The debater values the disagreement action similar to how someone playing a two armed bandit values the risky arm—namely, for more than its myopic expected payoff.

46. In Online Appendix Section B.3, we complement this analysis of one juror's *private* posterior over the course of the debate with an analysis of an outsider's *public* posterior.

We focus on the natural subgame. We define juror i's *dictatorial payoff* from his natural verdict[47] conditional on his own type y and his peer's type $x \geq \underline{x}$ as

$$P_i(\underline{x}, y) \equiv \int_{\underline{x}}^{\infty} \Delta_i(x-y) f(x|y, x \geq \underline{x}) dx. \tag{14}$$

A *devil's advocate* is a juror's type y in some period t of an equilibrium (x_t) who continues to argue for his natural verdict even though he would vote against it as dictator—formally, $y \geq x_t$ despite $P_i(x_{t-1}, y) < 0$. If $y \geq x_{t+2n}$, then type y is a devil's advocate in periods $t, t+2, \ldots, t+2n$. Indeed, we next argue that devil's advocacy may last many periods. We formulate this result expressing the delay cost as $\kappa_i = k_i \eta$, where $\eta > 0$ is the period length.

Proposition 8. (Devil's Advocacy). *Fix any $T^* > 0$ and $y^* \in \mathbb{R}$. For small biases $\beta_i \geq 0$ and period lengths $\eta > 0$, all types $y \geq y^*$ are devil's advocates for $T \geq T^*$ periods in any communicative equilibrium. Conversely, devil's advocacy is impossible if $\beta_L > 1 - 2\kappa_L$ and $\beta_M > 1 - 2\kappa_M$.*

If some type y of Lones infers in equilibrium that Moritz' (random) type x exceeds y, then a sufficiently unbiased Lones would strictly prefer to acquit as dictator. Then by continuity, there exists $\delta^* > 0$, such that the dictator y acquits conditional on the event that Moritz' (random) type x exceeds $y - \delta^*$. But as η vanishes, the hazard chance of debate ending vanishes, by the second claim in Proposition 3—and thus the interval of conceding types in any period vanishes too. So the number of periods t in which the dictator Lones is ready to acquit but the debater Lones plays devil's advocate, *i.e.* $x_t > y - \delta^*$ but $y > x_{t+1}$, is unbounded as η vanishes.[48]

Lones cannot be a devil's advocate when his bias or delay cost is large. In particular, if his miscarriage of justice decision costs is less than two periods delay cost, *i.e.* if $1 - \beta_L < 2\kappa_L$, then delay costs swamp the option value of conceding to strong types of Moritz in period $t+2$. So if the dictator Lones wishes to acquit in period t, then the debater Lones also concedes to acquittal.

Lones and Moritz may *simultaneously* play devil's advocates. Each may be among the weaker types yet to concede and so, as dictator, would throw in the towel, since each entertains the possibility that the other is strong—but debate continues to screen weak from strong types.

Devil's advocacy dynamically generalizes pivot voting, as pioneered by Austen-Smith and Banks (1996) and Feddersen and Pesendorfer (1996, 1997). Our devil's advocate persists in voting against his myopic best option because he conditions his vote on the weak conceding types of his peer. In these pivot voting papers, *sincere voting* is likewise not an equilibrium for like-minded but differentially informed voters. Rather, rational jurors should condition their vote on the event that it matters. These papers force pooling by all strong types for a position, whereas in our model, types separate by holding out. This benefit of deliberation is absent in the empirical investigation of appellate court decisions by Iaryczower *et al.* (forthcoming) because their model restricts attention to binary signals.

9. CONCLUSION

We develop and explore a new as-if model of debate. Notwithstanding our title, we find that the stereotypical win-lose acrimonious argument only emerges with low waiting costs or sufficiently

47. Devil's advocacy can also arise in the Nixon-China subgame. Indeed, with unbiased jurors the two subgames are identical; continuity thus suggests that they share qualitative features also for small biases. We omit a detailed analysis since it does not yield qualitatively different insights from the ones in the natural subgame.

48. This heuristic argument relies on the simplifying assumption that $x_{t+1} > x_t$, *i.e.* that cutoffs are ordered also between jurors, which need not be the case in equilibrium. The proof in Section A.12 dispenses with this assumption.

biased jurors. While such intransigent debate sees jurors eventually throwing in the towel on a debate, less biased or more impatient jurors engage in ambivalent debate, and eventually see a meeting of the minds. And when jurors are not too biased, greater polarization leads them to fight harder, and reach better informed verdicts.[49] In contrast, the basic insight emerging from the static cheap talk paper Crawford and Sobel (1982) is that greater bias hinders communication more. Our model also explains why playing devil's advocates is optimal.

For a useful link between our dynamic model and the existing static committee models, assume that our waiting costs vanish. While one might think that Li *et al*. (2001) corresponds to this zero cost limit, we underscore that LRS deduce ex post inefficient outcomes, where hawk and dove alike would wish the opposite verdict. But this torn conclusion of the debate only arises with our ambivalent debate, whereas small positive waiting costs in our model instead yields intransigent debate, which avoids such ex-post inefficiencies. Our paper also offers novel advice on how to select jurors to enhance information sharing and well-informed verdicts.

We conclude with a recent headline empirical finding: Gross *et al*. (2017) report that African-American prisoners who are convicted of murder are wrongfully convicted about 50% more often than other convicted murderers. Our theory opens the door to thinking more sharply about such juror bias studies. For instance, by Proposition 5, this emerges in intransigent debate if jurors are relatively biased against black defendants. In fact, the analysis in our paper jointly explains both debate errors *and* jury trial duration, and therefore offers hope of a stronger empirical link with the data; however, whereas our Proposition 6 explains how *symmetrically* increasing bias impacts debate duration—say a more hawkish hawk and dovish dove—in this racial context, we would instead require a more hawkish hawk and hawkish dove. Abandoning our symmetry assumptions is therefore an important direction for future research.

APPENDIX

A. OMITTED PROOFS

A.1. *Properties of the density functions*

The chance of state θ given type x obeys $\Gamma(\mathcal{G}|\ell)=1-\Gamma(\mathcal{I}|\ell)=\lambda=e^\ell/(1+e^\ell)$ and $\Gamma(\mathcal{I}|m)=1-\Gamma(\mathcal{G}|m)=\mu=e^m/(1+e^m)$. Then the density $f^\theta(x)$ of x in state $\theta=\mathcal{G},\mathcal{I}$ obeys $f^\mathcal{G}(\ell)/f^\mathcal{I}(\ell)=\Gamma(\mathcal{G}|\ell)/\Gamma(\mathcal{I}|\ell)=\lambda/(1-\lambda)=e^\ell$. The unconditional and conditional type densities are therefore

$$f(\ell)=\frac{1}{2}f^\mathcal{I}(\ell)+\frac{1}{2}f^\mathcal{G}(\ell)=\frac{1}{2}\left(1+\frac{f^\mathcal{G}(\ell)}{f^\mathcal{I}(\ell)}\right)f^\mathcal{I}(\ell)=\frac{1+e^\ell}{2}f^\mathcal{I}(\ell)=\frac{1+e^\ell}{2e^\ell}f^\mathcal{G}(\ell),$$

Then for any unconditional density f, there exist valid conditional signal densities f^θ. Next:

$$f(\ell|m)=f^\mathcal{I}(\ell)\Gamma(\mathcal{I}|m)+f^\mathcal{G}(\ell)\Gamma(\mathcal{G}|m)=\left(\frac{e^m}{1+e^m}+\frac{f^\mathcal{G}(\ell)/f^\mathcal{I}(\ell)}{1+e^m}\right)f^\mathcal{I}(\ell)=\frac{e^m+e^\ell}{1+e^m}f^\mathcal{I}(\ell),$$

Since $r(\ell,m)\equiv f(\ell|m)/f(\ell)$, the quotient of these expressions yields:

$$r(\ell,m)=\frac{f(\ell|m)}{f(\ell)}=\frac{2(e^\ell+e^m)}{(1+e^\ell)(1+e^m)}. \tag{15}$$

Then $r(\ell,m)$ is log-submodular, as it is a product of terms just in ℓ or m, and $e^\ell+e^m$—and if $a=e^\ell$ and $b=e^m$, then $(a'+b')(a+b)-(a'+b)(a+b')=-(a'-a)(b'-b)<0$ if $a'>a,b'>b$.

Assume ($\star$) in Section 3.5. We first relate the tails of the densities of signals $\phi(\lambda)$ and types $f(x)$.

Lemma A.1. (Signal Tails). *If the limit $\nu\equiv\lim_{\lambda\to0}\phi'(\lambda)\lambda/\phi(\lambda)$ exists and is finite, then the hazard rate of f is bounded with limit $\gamma^{-1}=\nu+1<\infty$.*

49. Bias also improves decisions in the advocacy models Dewatripont and Tirole (1999) and Che and Kartik (2009), where biased experts have stronger incentives to acquire information for their case.

Proof. Use the reverse type transformation $\lambda(x)=e^x/(1+e^x)$ with derivatives $\lambda'(x)=e^x/(1+e^x)^2$ and $\lambda''(x)=e^x(1-e^x)/(1+e^x)^3$, and the connection of the densities via $f(x)=\phi(\lambda(x))\lambda'(x)$. Then,

$$(\log f(x))'=\frac{\phi'(\lambda(x))\lambda'(x)}{\phi(\lambda(x))}+\frac{\lambda''(x)}{\lambda'(x)}=\frac{\phi'(\lambda(x))\lambda(x)}{\phi(\lambda(x))(1+e^x)}+\frac{1-e^x}{1+e^x}\to\nu+1$$

as $x\to-\infty$ and, by anti-symmetry, $\lim_{x\to\infty}(\log f(x))'=-\lim_{x\to-\infty}(\log f(x))'=-(\nu+1)$. So the likelihood ratio

$$\frac{f(x+a)}{f(x)}=\exp\left(\int_0^a(\log f(x+\hat{a}))'d\hat{a}\right)\to\exp(-(\nu+1)a) \tag{16}$$

as $x\to\infty$, and the inverse hazard rate converges to γ, as required:

$$\frac{\int_0^\infty f(x+a)da}{f(x)}=\int_0^\infty\exp\left(\int_0^a(\log f(x+\hat{a}))'d\hat{a}\right)da\to\int_0^\infty\exp(-(\nu+1)a)da=1/(\nu+1)=\gamma.$$

‖

We now analyse the distribution of the type difference $\delta=x-y$, conditional on own type y and the weakest remaining type of the opponent $y-\underline{\delta}$. Its density $f(y+\delta|y,\delta\geq-\underline{\delta})$ enters jurors' propensity function (9), which we analyse in detail in Section A.4. We show that this distribution of δ falls in y in the MLRP order, but converges to a limit distribution as $y\uparrow\infty$; we characterize this limit distribution explicitly and show that it is, surprisingly, *log-convex*.

Lemma A.2. (Density). *Fix $\underline{\delta}$ finite. Then (a) the conditional density $f(y+\delta|y,\delta\geq-\underline{\delta})$ is log-submodular in y,δ, (b) its limit $f^\infty(\delta|\delta\geq-\underline{\delta})\equiv\lim_{y\to\infty}f(y+\delta|y,\delta\geq-\underline{\delta})$ exists, and equals*

$$f^\infty(\delta|\delta\geq-\underline{\delta})=\frac{e^{-\delta/\gamma}(1+e^{-\delta})}{\int_{-\underline{\delta}}^\infty e^{-\hat{\delta}/\gamma}(1+e^{-\hat{\delta}a})d\hat{\delta}} \tag{17}$$

and (c) the limit density is log-convex.

Proof of Part (a). Decomposing $f(y+\delta|y,\delta\geq-\underline{\delta})=f(y+\delta)r(y+\delta,y)/(1-F(y-\underline{\delta}|y))$, we argue that the three RHS factors are log-submodular in y,δ. First, $f(y+\delta)$ is log-concave. Second, $1-F(y-\underline{\delta}|y)$ does not depend on δ and so is trivially (weakly) log-submodular in y,δ. Third,

$$r(y+\delta,y)=\frac{2(e^{y+\delta}+e^y)}{(1+e^{y+\delta})(1+e^y)}=\frac{2e^y(e^\delta+1)}{(1+e^ye^\delta)(1+e^y)}.$$

This is a product of log-submodular terms: $1/(1+e^ye^\delta)$ is log-submodular as $(1+ab)$ is log-supermodular: $(1+a'b')(1+ab)-(1+ab')(1+a'b)=(a'-a)(b'-b)>0$ if $a'>a,b'>b$. ‖

Proof of Part (b). Consider the likelihood ratio

$$\frac{f(y+\delta''|y,\delta\geq-\underline{\delta})}{f(y+\delta'|y,\delta\geq-\underline{\delta})}=\frac{f(y+\delta'')}{f(y+\delta')}\cdot\frac{r(y+\delta'',y)}{r(y+\delta',y)}. \tag{18}$$

The first quotient tends to $\exp(-(\delta''-\delta')/\gamma)$ as $y\uparrow\infty$ by (16). The second quotient

$$\frac{r(y+\delta'',y)}{r(y+\delta',y)}=\frac{e^y(1+e^{\delta''})}{(1+e^y)(1+e^{y+\delta''})}\cdot\frac{(1+e^y)(1+e^{y+\delta'})}{e^y(1+e^{\delta'})}=\frac{e^{-\delta''}+1}{e^{-\delta'}+1}\cdot\frac{e^{-\delta'}+e^y}{e^{-\delta''}+e^y} \tag{19}$$

tends to $(1+e^{-\delta''})/(1+e^{-\delta'})$ as $y\uparrow\infty$. Thus, for any δ',δ'', the likelihood ratio (18) converges to $e^{-\delta''/\gamma}(1+e^{-\delta''})/(e^{-\delta'/\gamma}(1+e^{-\delta'}))$. Then the density $f(y+\delta|y,\delta\geq-\underline{\delta})$, too, converges and its limit $f^\infty(\delta|\delta\geq-\underline{\delta})$ is proportional to $e^{-\delta/\gamma}(1+e^{-\delta})$. Scaling its integral to one yields (17). ‖

Proof of Part (c). Log-convexity of $f^\infty(\delta|\delta\geq-\underline{\delta})$ holds despite log-concavity of $f(y+\delta)$. For when $y\uparrow\infty$, the limiting unconditional $f(y+\delta)$ is exponential, proportional to $\exp(-\delta/\gamma)$, and log-convexity comes from correlation factor: $(\log(1+e^{-\delta}))'=-1/(1+e^\delta)$ rises in δ. ‖

A.2. *Monotone strategies: proof of Lemmas 1 and 2*

Proof of Lemma 1. Consider Moritz' choice to hold out until period t or t' in the natural subgame. If Lones concedes in period $s\in\{t+1,\cdots,t'-1\}$, then holding out until t' increases decision costs by $1-\beta_M$ if the state is $\mathcal{G}$, and reduces decision costs by $1+\beta_M$ if the state is $\mathcal{I}$. Also, holding out increases waiting costs by $(s-t)\kappa_M$. If Lones holds out past period $t'-1$, then Moritz's choice to hold out until period t' does not affect the verdict but increases waiting costs by

$(t'-t)\kappa_M$. Thus, when Lones' stopping time is $\varsigma(\ell)$, holding out until period t' increases the expected costs of Moritz's type m by:

$$\begin{aligned}&\Gamma(\mathcal{G}|m)\left[\sum_{s=t+1}^{t'-1}\int_{\varsigma^{-1}(s)}(1-\beta_M+(s-t)\kappa_M)f^{\mathcal{G}}(\ell)d\ell+\int_{\{\ell:\varsigma(\ell)>t'\}}(t'-t)\kappa_M f^{\mathcal{G}}(\ell)d\ell\right]+\\&\Gamma(\mathcal{I}|m)\left[\sum_{s=t+1}^{t'-1}\int_{\varsigma^{-1}(s)}(-(1+\beta_M)+(s-t)\kappa_M)f^{\mathcal{I}}(\ell)d\ell+\int_{\{\ell:\varsigma(\ell)>t'\}}(t'-t)\kappa_M f^{\mathcal{I}}(\ell)d\ell\right].\end{aligned}\tag{20}$$

The first line is positive—in state $\mathcal{G}$, if Moritz argues longer for $\mathcal{A}$, decision and waiting costs rise. But the second line has ambiguous sign: When Moritz holds out longer, decision costs fall but waiting costs rise. So if the costs (20) are negative, then the second line is negative. When m increases, costs (20) remain negative since $\partial_m\Gamma(\mathcal{I}|m)=-\partial_m\Gamma(\mathcal{G}|m)>0$. So we have proved a *single crossing property*: if m prefers to hold out from t to t', then so does any type $m'>m$.

So Moritz' best response in the natural subgame rise in his type; similar arguments apply to the Nixon-China subgame and to Lones' strategies in either subgame. ‖

Proof of Lemma 2. We show that Lones' best reply to a monotone agreeable strategy of Moritz is sincere, and conversely that Moritz' best reply to a monotone, sincere strategy of Lones is agreeable. For any agreeable monotone strategy of Moritz entails the ordered thresholds:

$$-\infty\leq\cdots\leq x_{-3}\leq x_{-1}<x_1\leq x_3\leq\cdots\leq\infty.$$

Let $t,t'\geq 2$ be even. Consider Lones' strategy $(\mathcal{C},t)$ to start natural debate and concede in period t, and his strategy $(\mathcal{A},t')$ to initiate Nixon-China debate and concede in period t'.

CLAIM. *The cost increment of* $(\mathcal{C},t)$ *over* $(\mathcal{A},t')$ *decreases in* ℓ.

Proof of Claim. The change $D(m)$ in waiting costs—$(\mathcal{C},t)$ less $(\mathcal{A},t')$—is monotone in m, since delay rises in m for natural debate, but falls in m for Nixon-China debate. Consider how decision costs change. If $m\in(x_{-t'-1},x_{t+1})$, the verdict changes, and the decision cost increment is $1-\beta_L$ in state $\mathcal{I}$ (more miscarriages of justice) and $-(1+\beta_L)$ in state $\mathcal{G}$ (fewer errors of impunity). The expected cost increment for Lones' type ℓ is:

$$\Gamma(\mathcal{I}|\ell)\left[\int_{x_{-t'-1}}^{x_{t+1}}(1-\beta_L)f^{\mathcal{I}}(m)dm+\int_{-\infty}^{\infty}D(m)f^{\mathcal{I}}(m)dm\right]$$

$$+\Gamma(\mathcal{G}|\ell)\left[\int_{x_{-t'-1}}^{x_{t+1}}-(1+\beta_L)f^{\mathcal{G}}(m)dm+\int_{-\infty}^{\infty}D(m)f^{\mathcal{G}}(m)dm\right].$$

Using $\partial_\ell\Gamma(\mathcal{G}|\ell)=-\partial_\ell\Gamma(\mathcal{I}|\ell)$, the derivative of the cost increment with respect to ℓ equals:

$$\partial_\ell\Gamma(\mathcal{G}|\ell)\left[\int_{x_{-t'-1}}^{x_{t+1}}((1-\beta_L)f^{\mathcal{I}}(m)+(1+\beta_L)f^{\mathcal{G}}(m))dm+\int_{-\infty}^{\infty}D(m)(f^{\mathcal{I}}(m)-f^{\mathcal{G}}(m))dm\right].$$

This is negative since $\Gamma(\mathcal{I}|\ell)=1/(1+e^{\ell})$ falls in ℓ, the first integral is positive, and the second integral is positive as $f^{\mathcal{I}}$ MLRP-dominates $f^{\mathcal{G}}$, as $f^{\mathcal{I}}(m)/f^{\mathcal{G}}(m)=e^m$. This proves the claim, and that Lones' best reply to an agreeable, monotone strategy is sincere. ‖

Conversely, consider Moritz' best response to a sincere strategy of Lones. If Lones is non-responsive, say he always initially argues $\mathcal{A}$, *i.e.* $x_0=\infty$, then—up to equivalence—we can set Moritz' first off-path cutoff x_1 equal to ∞, so that his strategy is agreeable. If Lones is responsive with finite initial cutoff x_0 and he initially argues $\mathcal{C}$, then by property $(P1)$ the cutoff gap x_1-x_0 must exceed $-\underline{b}_M$ to secure Moritz' indifference. Analogously, property $(\hat{P}1)$ guarantees $x_0-x_{-1}>\bar{b}_M$; hence $x_1-x_{-1}>\bar{b}_M-\underline{b}_M>0$, as desired. ‖

A.3. *Equilibrium characterization: Proof of Theorem 1*

First, consider sufficiency of the conditions. By definition, monotonicity (2) and $|x_0|<\infty$ imply that the strategy profile is sincere, agreeable, and responsive. Now suppose that cutoffs (x_t) are tight and obey indifference conditions (6) and (8), when finite.

We show that conceding in (odd) period $t+2$ of the natural subgame is optimal for any type $m\in[x_t,x_{t+2}]$ of Moritz. First, he weakly prefers this to stopping in period t because type $x_t\leq m$ is indifferent between these strategies, and the payoff difference quasi-increases in m, by Lemma 1. As $x_{t-2}\leq x_t\leq m$, the same argument shows that m weakly prefers conceding in period t to conceding in period $t-2$. By induction, ℓ does not want to concede before period $t+2$. Using this single-crossing logic, one can argue that concession period $t+2$ is weakly preferred to $t'>t+2$, *provided* $x_{t'}<\infty$. Finally, if $x_{t'}=\infty$ for some $t'>t$, then by tightness, no type of Lones concedes after period t', and so all types of Moritz prefer conceding in period t' to holding out longer—delay incurs waiting costs and does not change the verdict.

The analysis for Lones is similar, but for the initial period. For proposing $\mathcal{C}$ initially and conceding in the (even) period $t+2$ is a best response for Lones' types $\ell \in [x_t, x_{t+2}]$. We finish this argument, using the assumption that type x_0 is indifferent between initially voting $\mathcal{A}$ and $\mathcal{C}$ (conceding at once if Moritz disagrees) and the proof that the cost difference between these two plans is decreasing, by the Claim in the proof of Lemma 2.

Next consider necessity. We first argue that if no type of one juror concedes in period t then all types of the other juror concede in period $t-1$, that is, $x_{t-2} = x_t < \infty$ implies $x_{t-1} = \infty$. Suppose this fails, say, for even t. As no type of Lones concedes in period t, all types of Moritz prefer conceding in period $t-1$ over conceding in period $t+1$, implying $x_{t-1} = x_{t+1} < \infty$. As no type of Moritz concedes in period $t+1$, all types of Lones prefer conceding in period t over conceding in period $t+2$, implying $x_t = x_{t+2} < \infty$. Iterating this argument, we find that no type of either juror concedes after period $t-2$. But incurring infinite waiting costs with probability one is clearly incompatible with equilibrium.

So any sincere agreeable equilibrium is characterized by cutoffs (x_t) as in (2), where in each subgame there is a period t, possibly ∞, such that before t, the inequalities are strict, and in period $t+1$ the game almost surely ends. To fix ideas, assume WLOG even and finite t, so that $x_{t+1} = \infty$. Easily, the finite equilibrium cutoff types is indifferent: In any even period $t' < t$, in Lones' equilibrium strategy, types just below $x_{t'}$ concede in period t', and types just above $x_{t'}$ concede in period $t'+2$. By continuity of Lones' preferences in ℓ, the cutoff type $x_{t'}$ must be indifferent. The same argument shows that in any odd period $t' < t$, Moritz' equilibrium cutoff type $x_{t'}$ must be indifferent. Next, in periods $t' > t$ all types of Lones are indifferent between concession in periods t' and $t'+2$, because the game ends in period $t+1$ anyway. Finally, consider Lones' cutoff type x_t. As $x_{t+1} = \infty$ Lones is indifferent between conceding in period $t+2$ and conceding in any later period. In particular, it is a best reply for types just above x_t to concede in period $t+2$. By continuity, the equilibrium cutoff type x_t is indifferent between conceding and holding out in period t as required. Thus, (6) and (8) are necessary. ‖

A.4. *Properties of the propensity function*

Before proving the monotonicity properties $(P1)-(P5)$, we note analogous properties and one obvious additional property of the propensity to hold out in the Nixon-China subgame (5), as a function of own type y and cutoff gaps $\underline{\delta} \equiv y - \underline{x}$ and $\bar{\delta} \equiv \bar{x} - y$,

$$\hat{\pi}_i(\underline{\delta}, y, \bar{\delta}) \equiv \hat{\Pi}_i(y - \underline{\delta}, y, y + \bar{\delta}).$$

- $(\hat{P}1)$ *The propensity $\hat{\pi}_i$ quasi-increases in the upper gap $\bar{\delta}$ and is negative for all $\bar{\delta} < \bar{b}_i$*
- $(\hat{P}2)$ *The propensity $\hat{\pi}_i$ quasi-decreases in the type y,*
- $(\hat{P}3)$ *The propensity $\hat{\pi}_i$ is hump-shaped in the lower gap $\underline{\delta}$ with maximum at $-\underline{b}_i$,*
- $(\hat{P}4)$ *The propensity $\hat{\pi}_i$ decreases in the bias β_i and waiting cost κ_i.*
- $(\hat{P}5)$ *for y low enough, there exists $\epsilon > 0$ with $\partial\hat{\pi}_i/\partial\bar{\delta} > (1+\varepsilon)|\partial\hat{\pi}_i/\partial\underline{\delta}|$ when $\hat{\pi}_i(\underline{\delta}, y, \bar{\delta}) = 0$.*
- $(\hat{P}6)$ $\pi_i(\underline{\delta}, y, \bar{\delta}) - \hat{\pi}_i(\bar{\delta}, -y, \underline{\delta}) = 2\beta_i \geq 0$

Similarly, we write Lones' initial propensity to convict (7) as a function of cutoff gaps

$$\bar{\pi}_L(\underline{\delta}, y, \bar{\delta}) = \int_{-\infty}^{y-\underline{\delta}} \kappa_L f(y+\delta|y)d\delta + \int_{y-\underline{\delta}}^{y+\bar{\delta}} \Delta(\delta, \beta_L) f(y+\delta|y)d\delta - \int_{y+\bar{\delta}}^{\infty} \kappa_L f(y+\delta|y)d\delta.$$

- $(\bar{P}1)$ *The initial propensity $\bar{\pi}_L$ is U-shaped in the lower gap $\underline{\delta}$ with minimum at $\underline{\delta} = -\underline{b}_L$,*
- $(\bar{P}2)$ *The initial propensity $\bar{\pi}_L$ quasi-increases in the type y,*
- $(\bar{P}3)$ *The initial propensity $\bar{\pi}_L$ is hump-shaped in the upper gap $\bar{\delta}$, with maximum at $\bar{\delta} = \bar{b}_L$,*
- $(\bar{P}4)$ *The initial propensity $\bar{\pi}_L$ increases in the bias β_L.*

Properties $(P3)$, and $(P4)$ were proven in Section 3.5. To prove $(P1)$, we need to address the dependence of the distribution $f(y+\delta|y, \delta \geq -\underline{\delta})$ on $\underline{\delta}$. To do so, write (9) as

$$\frac{\int_{-\underline{\delta}}^{\bar{\delta}} \Delta(\delta, \beta_i) f(y+\delta|y)d\delta - \int_{\bar{\delta}}^{\infty} \kappa_i f(y+\delta|y)d\delta}{1 - F(y - \underline{\delta}|y)}.$$

The numerator quasi-increases in $\underline{\delta}$ as argued in Section 3.5, and the denominator is positive; hence, the fraction quasi-increases. To show $(P2)$ and the existence of the limit propensity $\pi_i^{\infty}(\underline{\delta}, \bar{\delta})$, note that the own type y affects the propensity

$$\pi_i(\underline{\delta}, y, \bar{\delta}) = \int_{-\underline{\delta}}^{\bar{\delta}} (\Delta(\delta, \beta_i) - \kappa_i) f(y+\delta|y, \delta \geq -\underline{\delta})d\delta - \int_{\bar{\delta}}^{\infty} 2\kappa_i f(y+\delta|y, \delta \geq -\underline{\delta})d\delta$$

only via the density $f(y+\delta|y, \delta \geq -\underline{\delta})$. By Lemma A.2(a), this δ distribution decreases in y in the MLRP, and so the integral of the quasi-decreasing integrand (falling on $[-\underline{\delta}, \bar{\delta}]$ and negative above $\bar{\delta}$) quasi-increases by Karlin and Rinott (1980),

Lemma 1. But $f(y+\delta|y,\delta\geq-\underline{\delta})$ converges as y increases, by Lemma A.2(b), and so the limit $\pi_i^{\infty}(\underline{\delta},\bar{\delta})\equiv\lim_{y\uparrow\infty}\pi_i(\underline{\delta},y,\bar{\delta})$ exists. Similarly, the *Nixon-China limit propensity* $\hat{\pi}_i^{\infty}(\underline{\delta},\bar{\delta})\equiv\lim_{y\downarrow-\infty}\hat{\pi}_i(\underline{\delta},y,\bar{\delta})$ exists, and inherits ($\hat{P}1$)–($\hat{P}5$).

We turn to ($P5$). Let $\phi(\delta)\equiv f(y+\delta|y)$ be the conditional type density, and Φ its survivor. Define the anti-symmetric $\psi(\delta)\equiv\frac{e^{\delta}-1}{e^{\delta}+1}$, so that $\Delta(\delta,\beta_i)=\psi(-\delta)+\beta_i$, and the *ex-ante propensity*

$$\check{\pi}_i(\underline{\delta},y,\bar{\delta})\equiv\Phi(-\underline{\delta})\pi_i(\underline{\delta},y,\bar{\delta})=\int_{-\underline{\delta}}^{\bar{\delta}}(\psi(-\delta)+\beta_i-\kappa_i)d\Phi(\delta)-2\kappa_i\int_{\bar{\delta}}^{\infty}d\Phi(\delta).$$

Since $\partial\check{\pi}_i/\partial\underline{\delta}=\Phi(-\underline{\delta})\partial\pi_i/\partial\underline{\delta}$ and $\partial\check{\pi}_i/\partial\bar{\delta}=\Phi(-\underline{\delta})\partial\pi_i/\partial\bar{\delta}$ whenever $\pi_i(\underline{\delta},y,\bar{\delta})=0$, property ($P5$) thus follows from its analogue for $\check{\pi}_i$, namely

$$\partial\check{\pi}_i/\partial\underline{\delta}>(1+\varepsilon)|\partial\check{\pi}_i/\partial\bar{\delta}|\quad\text{whenever}\quad\check{\pi}_i(\underline{\delta},y,\bar{\delta})=0. \tag{21}$$

Consider first intransigence, *i.e.* $\bar{\delta}<\bar{b}_i$, and fix y large enough. Lemma A.2(c) implies $\phi(-\underline{\delta})/(1-\Phi(-\underline{\delta}))<\phi(\bar{\delta})/(1-\Phi(\bar{\delta}))$. Since $\check{\pi}_i(\underline{\delta},y,\bar{\delta})=0$, delay costs $\kappa_i(2-\Phi(\bar{\delta})-\Phi(-\underline{\delta}))$ balance the net decision payoff, which is bounded above by $(\psi(-\underline{\delta})+\beta_i-\kappa_i)(\Phi(\bar{\delta})-\Phi(-\underline{\delta}))$ as ψ falls. Rearranging, $(\psi(-\underline{\delta})+\beta_i-\kappa_i)(1-\Phi(-\underline{\delta}))>(\psi(-\underline{\delta})+\beta_i+\kappa_i)(1-\Phi(\bar{\delta}))$. All told:

$$\frac{\partial\check{\pi}_i}{\partial\underline{\delta}}=(\psi(\underline{\delta})+\beta_i-\kappa_i)\phi(-\underline{\delta})\;>\;(\psi(\underline{\delta})+\beta_i-\kappa_i)\frac{1-\Phi(-\underline{\delta})}{1-\Phi(\bar{\delta})}\phi(\bar{\delta})>$$

$$>\;(\psi(\underline{\delta})+\beta_i+\kappa_i)\phi(\bar{\delta})>(\psi(-\bar{\delta})+\beta_i+\kappa_i)\phi(\bar{\delta})=\frac{\partial\check{\pi}_i}{\partial\bar{\delta}}.$$

The inequality $\psi(\underline{\delta})>(\psi(-\bar{\delta})$ has enough slack to cover the additional factor $1+\varepsilon$.

Consider ambivalence, *i.e.* $\bar{\delta}>\bar{b}_i$. Recalling (10), indifference balances the benefit of holding against weak types with the costs of holding out against strong types and the waiting costs

$$\mathcal{B}\equiv\int_{-\underline{b}_i}^{\underline{\delta}}(\psi(\delta)+\beta_i-\kappa_i)d\Phi(-\delta)\quad\text{and}\quad\mathcal{C}\equiv\int_{\underline{b}_i}^{\bar{\delta}}(\psi(\delta)-\beta_i+\kappa_i)d\Phi(\delta)+2\kappa_i\int_{\bar{\delta}}^{\infty}d\Phi(\delta). \tag{22}$$

The proof idea is that $\mathcal{B}$ is more sensitive to $\underline{\delta}$ than $\mathcal{C}$ is to $\bar{\delta}$, since the density ϕ is falling. Specifically, condition (21), and hence property ($P5$), follows from

$$\frac{\partial\check{\pi}_i/\partial\underline{\delta}}{\mathcal{B}}>(1+\varepsilon)\frac{-\partial\check{\pi}_i/\partial\bar{\delta}}{\mathcal{C}}. \tag{23}$$

To analyse these ratios, say the first, note that since ϕ decreases:

$$\frac{\partial\check{\pi}_i/\partial\underline{\delta}}{\mathcal{B}}=\frac{(\psi(\underline{\delta})+\beta_i-\kappa_i)\phi(-\underline{\delta})}{\int_{-\underline{b}_i}^{\underline{\delta}}(\psi(\delta)+\beta_i-\kappa_i)d\Phi(-\delta)}>\frac{\psi(\underline{\delta})+\beta_i-\kappa_i}{\int_{-\underline{b}_i}^{\underline{\delta}}(\psi(\delta)+\beta_i-\kappa_i)d\delta}. \tag{24}$$

So motivated, let $\Gamma(z)\equiv\int_{-\underline{b}_i}^{z}(\psi(\delta)+\beta_i-\kappa_i)d\delta$. The RHS of (24) then equals $\Gamma'(\underline{\delta})/\Gamma(\underline{\delta})$.

CASE 1 (SMALL BIASES). $\underline{b}_i=-|\underline{b}_i|<0$ *for small biases* $\beta_i<\kappa_i$, as depicted in Figure 8(a).

CLAIM 1. *If* $\underline{b}_i<0$, *then* $\Gamma(z)$ *is log-concave on* $[|\underline{b}_i|,\infty)$.

Proof. First, $\Gamma(|\underline{b}_i|)=\Gamma'(|\underline{b}_i|)=0$, and for any $z>|\underline{b}_i|$: $\Gamma(z)>0$ and $\Gamma'(z)=\psi(z)+\beta_i-\kappa_i>0$ and $\Gamma''(z)=\psi'(z)=2e^z/(1+e^z)^2>0$, and finally $\Gamma'''(z)=2e^z(1-e^z)/(1+e^z)^3<0$. Then

$$\Gamma(z)\;=\;\Gamma(|\underline{b}_i|)+\int_{|\underline{b}_i|}^{z}\Gamma'(\hat{z})d\hat{z}<(z-|\underline{b}_i|)\Gamma'(z)$$

$$\Gamma'(z)\;=\;\Gamma'(|\underline{b}_i|)+\int_{|\underline{b}_i|}^{z}\Gamma''(\hat{z})d\hat{z}>(z-|\underline{b}_i|)\Gamma''(z).$$

Claim 1 follows from $(\log\Gamma(z))''=[\Gamma(z)\Gamma''(z)-\Gamma'(z)^2]/\Gamma(z)^2<0$. ‖

We can assume $\underline{\delta}<\bar{\delta}$; otherwise, if $\underline{\delta}\geq\bar{\delta}$, then (23) follows by

$$\frac{\partial\check{\pi}_i/\partial\underline{\delta}}{-\partial\check{\pi}_i/\partial\bar{\delta}}=\frac{\psi(\underline{\delta})+\beta_i-\kappa_i}{\psi(\bar{\delta})-\beta_i-\kappa_i}\frac{\phi(-\underline{\delta})}{\phi(\bar{\delta})}>1+\varepsilon,$$

where $\varepsilon>0$ is the infimum over all $\phi(-\underline{\delta})/\phi(\bar{\delta})-1$ with $\check{\pi}_i(\underline{\delta},y,\bar{\delta})=0$.

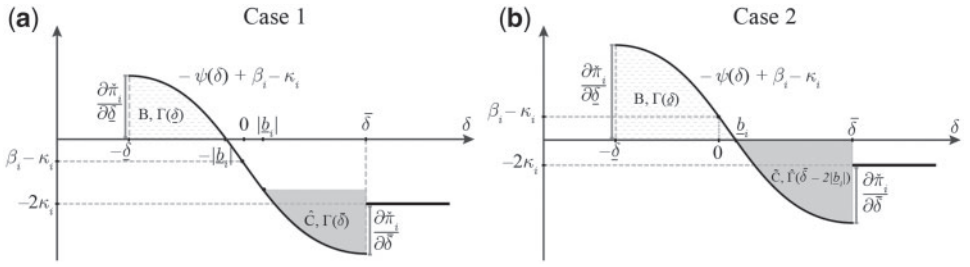

FIGURE 8
Analyzing the propensity. (a) Small bias, $\underline{b}_i<0$; (b) Large bias, $\underline{b}_i>0$

Dropping the second integral from (22) and shrinking the integrand and domain of the first integral, we define $\hat{\mathcal{C}}\equiv\int_{|\underline{b}_i|}^{\bar{\delta}}(\psi(\delta)+\beta_i-\kappa_i)d\Phi(\delta)$, as in Figure 8a. Since $\hat{\mathcal{C}}$ omits waiting costs for $\delta>\bar{\delta}$ and hazard rates are bounded, $\hat{\mathcal{C}}<\mathcal{C}/(1+\varepsilon)$ for some $\varepsilon>0$. Now

$$\frac{\partial\check{\pi}_i/\partial\underline{\delta}}{\mathcal{B}}=\frac{(\psi(\underline{\delta})+\beta_i-\kappa_i)\phi(-\underline{\delta})}{\int_{|\underline{b}_i|}^{\underline{\delta}}(\psi(\delta)+\beta_i-\kappa_i)d\Phi(-\delta)}>\frac{\psi(\underline{\delta})+\beta_i-\kappa_i}{\int_{|\underline{b}_i|}^{\underline{\delta}}(\psi(\delta)+\beta_i-\kappa_i)d\delta}=\frac{\Gamma'(\underline{\delta})}{\Gamma(\underline{\delta})} \tag{25}$$

$$\frac{-\partial\check{\pi}_i/\partial\bar{\delta}}{\hat{\mathcal{C}}}=\frac{(\psi(\bar{\delta})-\beta_i-\kappa_i)\phi(\bar{\delta})}{\int_{|\underline{b}_i|}^{\bar{\delta}}(\psi(\delta)+\beta_i-\kappa_i)d\Phi(\delta)}<\frac{\psi(\bar{\delta})+\beta_i-\kappa_i}{\int_{|\underline{b}_i|}^{\bar{\delta}}(\psi(\delta)+\beta_i-\kappa_i)d\delta}=\frac{\Gamma'(\bar{\delta})}{\Gamma(\bar{\delta})} \tag{26}$$

so that (23) follows from (using $\underline{\delta}<\bar{\delta}$ and Claim 1):

$$\frac{\partial\check{\pi}_i/\partial\underline{\delta}}{\mathcal{B}}>\frac{\Gamma'(\underline{\delta})}{\Gamma(\underline{\delta})}>\frac{\Gamma'(\bar{\delta})}{\Gamma(\bar{\delta})}>\frac{-\partial\check{\pi}_i/\partial\bar{\delta}}{\hat{\mathcal{C}}}>(1+\varepsilon)\frac{-\partial\check{\pi}_i/\partial\bar{\delta}}{\mathcal{C}}.$$

CASE 2 (LARGE BIAS). $\underline{b}_i>0$ *for large biases* $\beta_i>\kappa_i$, as depicted in Figure 8b.

CLAIM 2. If $\underline{b}_i>0$, then $\hat{\Gamma}(z)\equiv\int_{\underline{b}_i}^{z+2\underline{b}_i}(\psi(\delta)-\beta_i+\kappa_i)d\delta$ is log-concave and for all $z>-\underline{b}_i$:

$$\Gamma'(z)\hat{\Gamma}(z)-\Gamma(z)\hat{\Gamma}'(z)>0\qquad\text{and}\qquad\Gamma''(z)\hat{\Gamma}'(z)-\Gamma'(z)\hat{\Gamma}''(z)>0. \tag{27}$$

Proof. Claim 1 gives log-concavity. We first prove (27) for z near $-\underline{b}_i$: Now, $\Gamma(-\underline{b}_i)=\Gamma'(-\underline{b}_i)=\hat{\Gamma}(-\underline{b}_i)=\hat{\Gamma}'(-\underline{b}_i)=0$, so that (27) hold with equality for $z=-\underline{b}_i$. Also, $\Gamma''(-\underline{b}_i)=\psi'(-\underline{b}_i)=\psi'(\underline{b}_i)=\hat{\Gamma}''(-\underline{b}_i)$, and $\Gamma'''(-\underline{b}_i)=\psi''(-\underline{b}_i)>0>\psi''(\underline{b}_i)=\hat{\Gamma}'''(-\underline{b}_i)$. Then the first four terms of the Taylor expansion of (27) around $z=-\underline{b}_i$ vanish, but the fifth is positive: $2(\Gamma'''(-\underline{b}_i)\hat{\Gamma}''(-\underline{b}_i)-\Gamma(-\underline{b}_i)''\hat{\Gamma}'''(-\underline{b}_i))=2(\Gamma'''(-\underline{b}_i)-\hat{\Gamma}'''(-\underline{b}_i))\psi'(-\underline{b}_i)$. Similarly, the first two terms in a Taylor expansion of the second inequality in (27) around $z=-\underline{b}_i$ vanish, but the third is positive.

To conclude, we show that both left sides in (27) quasi-increase. Indeed, $\psi'(z)=2e^z/(1+e^z)^2$, $\psi''(z)=2e^z(1-e^z)/(1+e^z)^3$ and hence $\psi''(z)/\psi'(z)=(1-e^z)/(1+e^z)$ falls in z. Thus,

$$\frac{\Gamma'''(z)}{\Gamma''(z)}=\frac{\psi''(z)}{\psi'(z)}>\frac{\psi''(z+2\underline{b}_i)}{\psi'(z+2\underline{b}_i)}=\frac{\hat{\Gamma}'''(z)}{\hat{\Gamma}''(z)}.$$

So, if the second inequality in (27) holds with equality at z, then

$$0=\Gamma''(z)\hat{\Gamma}'(z)-\Gamma'(z)\hat{\Gamma}''(z)<\Gamma'''(z)\hat{\Gamma}'(z)-\Gamma'(z)\hat{\Gamma}'''(z)=\left(\Gamma''(z)\hat{\Gamma}'(z)-\Gamma'(z)\hat{\Gamma}''(z)\right)',$$

and it holds for all $z>-\underline{b}_i$. To show that the first inequality holds for $z>-\underline{b}_i$, we repeat the last step: Assume (27) held with equality at some z. Then, by the second inequality in (27):

$$0=\Gamma'(z)\hat{\Gamma}(z)-\Gamma(z)\hat{\Gamma}'(z)<\Gamma''(z)\hat{\Gamma}(z)-\Gamma(z)\hat{\Gamma}''(z)=\left(\Gamma'(z)\hat{\Gamma}(z)-\Gamma(z)\hat{\Gamma}'(z)\right)'.$$

‖

We can assume $\bar{\delta}-2\underline{b}_i>\underline{\delta}$: Otherwise $\psi(\bar{\delta})-\psi(\underline{\delta})=\int_{\underline{\delta}}^{\bar{\delta}}\psi'(\delta)d\delta<\int_{-\underline{b}_i}^{\underline{b}_i}\psi'(\delta)d\delta=2\psi(\underline{b}_i)=2(\beta_i-\kappa_i)<2\beta_i$ and thus (23) follows by

$$\frac{\partial\check{\pi}_i/\partial\underline{\delta}}{-\partial\check{\pi}_i/\partial\bar{\delta}}=\frac{\psi(\underline{\delta})+\beta_i-\kappa_i}{\psi(\bar{\delta})-\beta_i-\kappa_i}\frac{\phi(-\underline{\delta})}{\phi(\bar{\delta})}>1+\varepsilon.$$

In analogy to the set $\hat{\mathcal{C}}$ in case 1, define $\tilde{\mathcal{C}} \equiv \int_{\underline{b}_i}^{\overline{\delta}} (\psi(\delta) - \beta_i + \kappa_i) d\Phi(\delta)$, as in Figure 8b. Again, we clearly have $\tilde{\mathcal{C}} < \mathcal{C}/(1+\varepsilon)$ for some $\varepsilon > 0$. Also, as in (26),

$$\frac{-\partial \breve{\pi}_i / \partial \overline{\delta}}{\tilde{\mathcal{C}}} = \frac{(\psi(\overline{\delta}) - \beta_i - \kappa_i)\phi(\overline{\delta})}{\int_{\underline{b}_i}^{\overline{\delta}} (\psi(\delta) - \beta_i + \kappa_i) d\Phi(\delta)} < \frac{\psi(\overline{\delta}) - \beta_i + \kappa_i}{\int_{\underline{b}_i}^{\overline{\delta}} (\psi(\delta) - \beta_i + \kappa_i) d\delta} = \frac{\hat{\Gamma}'(\overline{\delta} - 2\underline{b}_i)}{\hat{\Gamma}(\overline{\delta} - 2\underline{b}_i)}. \tag{28}$$

Exploiting (25), (27), Claim 2 and $\overline{\delta} - 2\underline{b}_i > \underline{\delta}$, (28), and $\tilde{\mathcal{C}} < \mathcal{C}/(1+\varepsilon)$, respectively:

$$\frac{\partial \breve{\pi}_i / \partial \underline{\delta}}{\mathcal{B}} > \frac{\Gamma'(\underline{\delta})}{\Gamma(\underline{\delta})} > \frac{\hat{\Gamma}'(\underline{\delta})}{\hat{\Gamma}(\underline{\delta})} > \frac{\hat{\Gamma}'(\overline{\delta} - 2\underline{b}_i)}{\hat{\Gamma}(\overline{\delta} - 2\underline{b}_i)} > \frac{-\partial \breve{\pi}_i / \partial \overline{\delta}}{\tilde{\mathcal{C}}} > (1+\varepsilon) \frac{-\partial \breve{\pi}_i / \partial \overline{\delta}}{\mathcal{C}}.$$

This concludes the proof of (P5).

A.5. *Equilibrium existence: Proof of Theorem 2*

PROOF PLAN. In the $(\sigma, \tau) = (1,1)$-equilibrium, all Moritz' types second Lones' proposal, and existence requires $\bar{\Pi}_L(-\infty, x_0, \infty) = 0$, for some x_0. Lones' conviction propensity $\bar{\Pi}_L(-\infty, x_0, \infty)$ is continuous in x_0, negative for low x_0, when Lones thinks the defendant innocent, and positive for high x_0, when he thinks him guilty. Indifference holds at $x_0 = \log((1-\beta_L)/(1+\beta_L))$.

More generally, we must solve for the multi-dimensional (and possibly infinite-dimensional) cutoff vector $x_{-\sigma+1}, \ldots, x_{\tau-1}$. This requires a multi-dimensional extension of the intermediate value theorem. Specifically, define for every finite $k \geq 2$ the *iterated domain* $X(k)$ as the set of cutoff pairs $(x_0, x_1) \in \mathbb{R}^2$ compatible with deferential natural debate with drop-dead date k, *i.e.* there are finite cutoffs $(x_t)_{t=2,\ldots,k-1}$ with for $t = 1, \ldots, k-1$:

$$\Pi_{i(t)}(x_{t-1}, x_t, x_{t+1}) = 0, \tag{29}$$

where $x_k = \infty$; for $k = 1$, we require $x_1 = \infty$ and so set $X(1) \equiv \{(x_0, \infty) : x_0 \in \mathbb{R}\}$. Similarly, $X(\infty)$ is the set of cutoff pairs (x_0, x_1) compatible with natural communicative debate, *i.e.* with finite cutoffs $(x_t)_{t>1}$ obeying (29) for all $t > 1$. Analogously, $\hat{X}(k)$ is the sets of cutoff pairs (x_0, x_{-1}) in the Nixon-China subgame compatible with deferential/communicative debate. By Theorem 1, a (σ, τ)-equilibrium exists iff there are (x_{-1}, x_0, x_1) with $(x_0, x_1) \in X(\tau)$, $(x_0, x_{-1}) \in \hat{X}(\sigma)$, and $\bar{\Pi}_L(x_{-1}, x_0, x_1) = 0$.

MATHEMATICAL PRELIMINARIES. First, we consider types ℓ and m as elements of the (compact) extended reals $\bar{\mathbb{R}} \equiv \mathbb{R} \cup \{\pm\infty\}$. Let $Y(k) \subset \bar{\mathbb{R}}^2$ be the topological closure of $X(k)$, and similarly $\hat{Y}(k)$ for the closure of $\hat{X}(k)$. Second, for sets $X, Y, Z \subseteq \bar{\mathbb{R}}^2$, say X *connects* Y *and* Z if X has a connected component that intersects both Y and Z. Then X does not admit an open, disjoint cover Ω_1 and Ω_2, where $\Omega_1 \cap Y = \emptyset$ and $\Omega_2 \cap Z = \emptyset$.

We then prove existence by first showing in Lemma A.3 that $Y(k)$ connects the *left edge* $\mathcal{L} \equiv \{-\infty\} \times \mathbb{R}$ of $\bar{\mathbb{R}}^2$ to its *upper right corner* $\mathcal{UR} \equiv (\infty, \infty)$, and $\hat{Y}(\hat{k})$ connects the *right edge* $\mathcal{R} \equiv \{\infty\} \times \mathbb{R}$ to the *lower left corner* $\mathcal{DL} \equiv (-\infty, -\infty)$. Then in Lemma A.4 there are (x_{-1}, x_0, x_1) with $(x_0, x_1) \in X(\tau)$, $(x_0, x_{-1}) \in \hat{X}(\sigma)$, and $\bar{\Pi}_L(x_{-1}, x_0, x_1) = 0$. Figure 9(a) depicts these elements and subsets of $\bar{\mathbb{R}}^2$, and is an existence proof guide.

THE ITERATED DOMAINS. Write cutoffs x_{t-1} backwards, as a function of successive cutoffs x_t, x_{t+1} via the indifference condition (29). By (*P*1) there is at most one root $x_{t-1} < x_{t+1}$. So motivated, define the *backward shooting function* ξ_i and its domain $\mathcal{R}_i$ as all $(y, \bar{x}) \in \bar{\mathbb{R}}^2$ such that $\Pi_i(\cdot, y, \bar{x}) = 0$ admits a (unique) root $\xi_i(y, \bar{x}) < \bar{x}$ (possibly equal to $-\infty$). By convention, $\mathcal{UR} \in \mathcal{R}_i$ and $\xi_i(\mathcal{UR}) \equiv \infty$. For convenience, also define $\Xi_i : \mathcal{R}_i \to \bar{\mathbb{R}}^2$ by $\Xi_i(y, \bar{x}) \equiv (\xi_i(y, \bar{x}), y)$.

Define $\Xi_i(X) = \Xi_i(X \cap \mathcal{R}_i)$ whenever $X \not\subseteq \mathcal{R}_i$. Let $Y(1) = \{(x_0, \infty) : x_0 \in \bar{\mathbb{R}}\}$ and, inductively $Y(2k) = \Xi_M \circ (\Xi_L \circ \Xi_M)^{k-1}(Y(1))$ and $Y(2k+1) = (\Xi_M \circ \Xi_L)^k(Y(1))$ for $k > 1$. Then $X(k) = Y(k) \cap \mathbb{R}^2$, since both sets are defined by (29), but cutoffs are finite for $X(k)$. For $k = \infty$, define the sets $Z(1) \equiv \bar{\mathbb{R}}^2$, $Z(2k+1) \equiv (\Xi_M \circ \Xi_L)^k(Z(1))$ and $Y(\infty) \equiv \bigcap_k Z(2k+1)$. Clearly $Z(2k+1) \subseteq Z(2k-1)$ for any k, and $X(\infty) = Y(\infty) \cap \mathbb{R}^2$. In the Nixon-China subgame, we define the analogous sets $\hat{Y}(k)$ with $\hat{X}(k) = \hat{Y}(k) \cap \mathbb{R}^2$.

Lemma A.3. *For any k, $Y(k)$ connects $\mathcal{UR}$ and $\mathcal{L}$, and $\hat{Y}(k)$ connects $\mathcal{DL}$ and $\mathcal{R}$.*

STEP 1. *$\mathcal{UR} \in \mathcal{R}_i$, $\mathcal{L} \cap \mathcal{R}_i = \emptyset$, and $\partial \mathcal{R}_i = \Xi_i^{-1}(\mathcal{L})$, i.e., the topological boundary of $\mathcal{R}_i$.*

Proof. $\mathcal{UR} \in \mathcal{R}_i$ follows from $\Xi_i(\mathcal{UR}) = \mathcal{UR}$. For the other properties of $\mathcal{R}_i$, recall that $\Pi_i(\underline{x}, y, \bar{x})$ quasi-decreases in $\underline{x}$ by (P1), and is negative as $\underline{x}$ tend to $\bar{x}$, by (4). Thus, $\mathcal{R}_i = \{(y, \bar{x}) : \Pi_i(-\infty, y, \bar{x}) \geq 0\}$ and compactness follows by continuity of Π_i. The last two assertions of Step 1 then follow since $\Pi_i(-\infty, y, \bar{x}) < 0$ for low y, and quasi-increases in y by Lemma 1. ‖

STEP 2. *$\xi_i : \mathcal{R}_i \to \bar{\mathbb{R}}^2$ is continuous.*

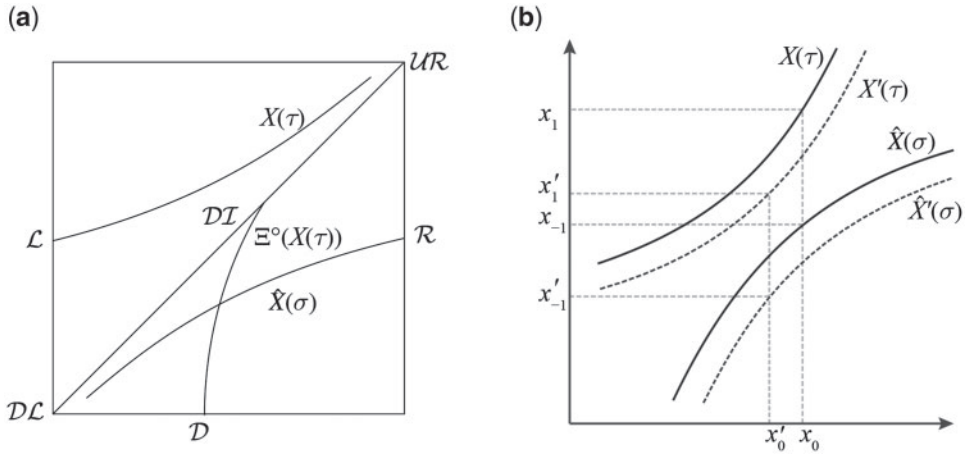

FIGURE 9

Equilibrium existence and comparative statics (ambivalence). Panel (a) depicts the existence proof logic in Section A.5, and panel (b) the proof logic for Proposition 4 in Section A.10

Proof. For $(y,\bar{x})\in\mathcal{R}_i\setminus\{\mathcal{UR}\}$ this follows because $\Pi_i(\underline{x},y,\bar{x})$ is continuous and (strictly) increases in $\underline{x}$ at $\xi_i(y,\bar{x})$. At $(y,\bar{x})=\mathcal{UR}$, by definition $\xi_i(\mathcal{UR})=\infty$, so to prove continuity, we need that when $y,\bar{x}$ are large, $\xi_i(y,\bar{x})$ is also large: As $\bar{x}\to\infty$ the second integral in (4) vanishes; and as $y\to\infty$, the integrand of the first integral tends to $1+\beta_i$ on almost all of its domain (where $\Delta(x-y,\beta_i)$ tends to $1+\beta_i-\kappa_i$). Thus, for fixed $\underline{x}$, the propensity π_i approximates $1+\beta_i-\kappa_i>0$; to restore indifference, we must have $\underline{x}\to\infty$, too. ‖

STEP 3. *If X connects $\mathcal{UR}$ and $\mathcal{L}$ then so does $\Xi_i(X)$.*

Proof. Assume to the contrary that $\Xi_i(X)$ is covered by two disjoint open sets Ω_1 and Ω_2, where $\Omega_1\cap\mathcal{L}=\emptyset$ and $\mathcal{UR}\notin\Omega_2$. By continuity (shown in Step 2), $\Pi_1\equiv\Xi_i^{-1}(\Omega_1)$ is open in $\mathcal{R}_i$. As $\Omega_1\cap\mathcal{L}=\emptyset$ and $\Xi_i^{-1}(\mathcal{L})=\partial\mathcal{R}_i$ (shown in Step 1), Π_1 lies in the interior of $\mathcal{R}_i$, and so also open in $\bar{\mathbb{R}}^2$. As $\Pi_1\subset\mathcal{R}_i$ and $\mathcal{R}_i\cap\mathcal{L}=\emptyset$, we have $\Pi_1\cap\mathcal{L}=\emptyset$. By continuity, $\Xi_i^{-1}(\Omega_2)$ is open in $\mathcal{R}_i$ and so $\Pi_2\equiv\Xi_i^{-1}(\Omega_2)\cup(\bar{\mathbb{R}}^2\setminus\mathcal{R}_i)$ is open in $\bar{\mathbb{R}}^2$. As $\mathcal{UR}\in\mathcal{R}_i$ (shown in Step 1), $\mathcal{UR}\notin\Omega_2$ by definition of Ω_2, and so $\mathcal{UR}\notin\Xi_i^{-1}(\Omega_2)$, we get $\mathcal{UR}\notin\Pi_2$. Then Π_1 and Π_2 is an open cover of X with $\Pi_1\cap\mathcal{L}=\emptyset$ and $\mathcal{UR}\notin\Pi_2$, contrary to the premise of Step 3. ‖

Proof of Lemma A.3. By definition, $Y(1)=\{(x_0,\infty):x_0\in\bar{\mathbb{R}}\}$ connects $\mathcal{UR}$ and $\mathcal{L}$. Step 3 then implies inductively that $Y(k)$ connects $\mathcal{UR}$ and $\mathcal{L}$ for every finite k. We prove Lemma A.3 for $Y(\infty)$. By construction, $Z(2k+1)\supseteq Y(2k+1)$ and thus $Z(2k+1)$ connects $\mathcal{UR}$ and $\mathcal{L}$. Also, the sets $Z(2k+1)$ are compact and 'decreasing', *i.e.* $Z(2k+1)\subseteq Z(2k-1)$. By Cantor's intersection theorem, $Y(\infty)=\bigcap_k Z(2k+1)\neq\varnothing$ is compact.

We now argue that $Y(\infty)$ connects $\mathcal{UR}$ and $\mathcal{L}$. If not, it is openly covered by Ω_1,Ω_2 where $\Omega_1\cap\mathcal{L}=\emptyset$ and $\mathcal{UR}\notin\Omega_2$. Let $Y_k\equiv\bar{\mathbb{R}}^2\setminus Z(2k+1)$ $\{\Omega_1,\Omega_2,(Y_k)_{k\geq1}\}$ be an open cover of the compact space $\bar{\mathbb{R}}^2$, with a finite open subcover $\{\Omega_1,\Omega_2,Y_k\}$ for some finite k. Then Ω_1,Ω_2 is an open cover of $Z(2k+1)$, which contradicts the fact that $Z(2k+1)$ connects $\mathcal{UR}$ and $\mathcal{L}$.

This establishes Lemma A.3 for $Y(k)$; the proof for $\hat{Y}(k)$ is analogous. ‖

THE INITIAL PERIOD. To establish existence of the deferential equilibrium with drop-dead dates (σ,τ), choose finite (x_{-1},x_0,x_1) with $(x_0,x_1)\in X(\tau)$, $(x_0,x_{-1})\in\hat{X}(\sigma)$ and $\bar{\Pi}_L(x_{-1},x_0,x_1)=0$. As above, it is useful to write x_{-1} backwards, as a function of x_0,x_1 via x_0's indifference condition (8). As $\bar{\Pi}_L$ is U-shaped in x_{-1} with minimum at $x_0+\underline{b}_L$ by $(\bar{P}1)$, there can be at most one root $x_{-1}<x_0-\bar{b}_L$ (possibly $-\infty$). Such a root exists iff $\bar{\Pi}_L(-\infty,x,y)\geq0\geq\bar{\Pi}_L(x-\bar{b}_L,x,y)$; if it exists we denote it by $\bar{\xi}(x,y)$. By the proof of Step 2 in the proof of Lemma A.3, $\bar{\xi}$ is continuous on its domain. Also define $\bar{\Xi}(x,y)\equiv(x,\bar{\xi}(x,y))$. Existence then follows from:

Lemma A.4. *For any σ,τ, the sets $\bar{\Xi}(X(\tau))$ and $\hat{X}(\sigma)$ intersect.*

Proof. First, (7) implies $\bar{\Pi}_L(-\infty,x,y)<0$ for $(x,y)\in\mathcal{L}$, *i.e.* $x=-\infty$ since then Lones knows that the defendant is guilty and that Moritz will second a convict proposal. Also, by $(\bar{P}1)$ in Section A.4, we have $\bar{\Pi}_L(x-\bar{b}_L,x,y)<\bar{\Pi}_L(-\infty,x,y)$ for all finite (x,y).

As $Y(\tau)$ connects $\mathcal{L}$ to $\mathcal{UR}$, by Lemma A.3, and $X(\tau)=\bar{\mathbb{R}}^2\cap Y(\tau)$, we see that $X(\tau)$ connects $\{(x,y):\bar{\Pi}_L(x-\bar{b}_L,x,y)=0\}$ and $\{(x,y):\bar{\Pi}_L(-\infty,x,y)=0\}$. By continuity, $\bar{\Xi}(X(\tau))$ connects the upper diagonal $\mathcal{DI}\equiv\{(x,x-\bar{b}_L):x\in\mathbb{R}\}$ and the lower edge $\mathcal{D}\equiv\{(x,-\infty):x\in\mathbb{R}\}$ of the "lower triangle" $\mathcal{LT}\equiv\{(x,y)\in\bar{\mathbb{R}}^2:y\leq x-\bar{b}_L\}$.

Graphically, $\bar{\Xi}(X(\tau))$ and $\hat{X}(\sigma)$ intersect, by Figure 9a. For if not, $\bar{\Xi}(X(\tau))$ belongs to one of the (open) connected components X of $\mathcal{LT}\setminus\hat{X}(\sigma)$. As an open and connected set, X is path-connected; by set-inclusion, it connects $\mathcal{DI}$ and $\mathcal{D}$, and by construction, it does not intersect $\hat{X}(\sigma)$. By the Jordan curve theorem, X divides $\mathcal{LT}$ into two connected components—one includes $\mathcal{DL}$, and the other $\mathcal{R}$. But $\hat{X}(\sigma)$ connects $\mathcal{DL}$ and $\mathcal{R}$. Contradiction. ‖

A.6. *Stability: Proof of Theorem 4*

Fix an equilibrium where Lones, say, is expected to defer in period τ. Not *all* remaining types of Moritz need not switch in period $\tau+1$, but *enough of them* do that Lones' strongest remaining types prefer to concede. So defining $\bar{m}$ as the solution of $\Pi_L(x_{\tau-1},\infty,\bar{m})=0$, Lones' deference requires some (weak) type $m\leq\bar{m}$ of Moritz to hold out in period τ.

In turn, forward induction requires Moritz to attribute an unexpected deviation by Lones to (strong) types ℓ who could benefit from it given some continuation strategy of Moritz, that is types above the threshold $\tilde{\ell}$ which solves $\Pi_L(x_{\tau-1},\tilde{\ell},\tilde{\ell}+\bar{b}_L)=0$. Since Π_L quasi-increases in own type by Lemma 1, and $\Pi_L(x_{\tau-1},\infty,\cdot)$ increases, the definitions of $\bar{m}$ and $\tilde{\ell}$ imply

$$\bar{m}<\tilde{\ell}+\bar{b}_L-\varepsilon \tag{30}$$

for some $\varepsilon>0$ that is independent of κ_L,β_L and $x_{\tau-1}$, because $\tilde{\ell}-x_{\tau-1}$ is boundedly finite. Hence, $\Pi_L(x_{\tau-1},\infty,\tilde{\ell}+\bar{b}_L)-\Pi_L(x_{\tau-1},\tilde{\ell},\tilde{\ell}+\bar{b}_L)$ is bounded below.

From (10), given Lones' type ℓ, Moritz' type m prefers acquittal tomorrow to conviction today iff $\ell<m+\underline{b}_M$. So the equilibrium is unstable if $\tilde{\ell}>\bar{m}+\underline{b}_M$, *i.e.* if Lones' deviating type is strong given Moritz' insisting type, but is equivalent to a stable equilibrium if $\tilde{\ell}<\bar{m}+\underline{b}_M$.

Now, assume small β_L,β_M and $\kappa_L=\kappa_M$. The thresholds $-\underline{b}_M$ and $\bar{b}_L$ in (10) are then close to each other. So inequality (30) implies $\tilde{\ell}>\bar{m}+\underline{b}_M$, and the equilibrium is unstable.

For a deferential equilibrium with drop-dead date σ for Nixon-China debate, define Lones' weakest (highest) type $\hat{\ell}$ for whom Moritz can rationalize a deviation in period σ, and Moritz' weakest (highest) type $\hat{m}$ that holds out in period $\sigma+1$ to ensure Lones' deference. We must then have $\hat{m}>\hat{\ell}+\underline{b}_L$, and a fortiori $\hat{\ell}<\hat{m}+\bar{b}_M$. With this condition, Moritz' type $\hat{m}$ facing Lones' type $\hat{\ell}$ prefers an acquittal today over a conviction tomorrow. The equilibrium is unstable. ‖

A.7. *Ambivalence: Proof of Proposition 1*

We first argue that the assumptions of Proposition 1 imply (12). Since the propensity $\pi_i(\underline{\delta},y,\bar{\delta})$ quasi-increases in y by property $(P2)$ and is maximized for $\bar{\delta}=\bar{b}_i$ by property $(P3)$, inequalities (12) are equivalent to $\pi_M^\infty(\bar{b}_L,\bar{b}_M)<0$ and $\pi_L^\infty(\bar{b}_M,\bar{b}_L)<0$, for the asymptotic propensity

$$\pi_i^\infty(\underline{\delta},\bar{\delta})=\int_{-\underline{\delta}}^{\bar{\delta}}(\Delta(\delta,\beta_i)-\kappa_i)f^\infty(\delta|\delta\geq-\underline{\delta})d\delta-\int_{\bar{\delta}}^{\infty}2\kappa_i f^\infty(\delta|\delta\geq-\underline{\delta})d\delta$$

from Section 3.5, and the asymptotic density $f^\infty(\delta|\delta\geq-\underline{\delta})$ from Lemma A.2(b).

If preferences are common, then $\bar{b}_L=\bar{b}_M=-\underline{b}_L=-\underline{b}_M$, and so $(P1)$ implies $\pi_M^\infty(\bar{b}_L,\bar{b}_M)<0$ and $\pi_L^\infty(\bar{b}_M,\bar{b}_L)<0$. By continuity, these two inequalities hold for closely aligned interests.

Next, for sufficiently informative types, the type density $f^\infty(\delta|\delta\geq-\underline{\delta})$ in (17) assigns most probability to $\delta>\bar{b}_i$, and $\pi_M^\infty(\bar{b}_L,\bar{b}_M)$, say, converges to $-2\kappa_M<0$ as $\gamma\uparrow\infty$.

To see that (12) implies ambivalence, assume to the contrary that, say, Moritz is intransigent in period t, *i.e.* $\delta_t<\bar{b}_M$. Then $\pi_L(\delta_t,x_t,\bar{\delta})<0$ by the second inequality in (12) and property $(P1)$. This contradiction to x_t's indifference condition implies ambivalence. ‖

A.8. *Ambivalent Nixon-China Debate: proof of Proposition 2*

To see that $|\kappa_M-\kappa_L|\leq\beta_L+\beta_M$ ensures ambivalent Nixon-China debates, assume to the contrary that, say, Moritz is intransigent in period $t-1$. Given (5), this is equivalent to $\delta_{-(t-1)}\leq-\underline{b}_M$. Given (10) and $\kappa_M-\kappa_L\leq\beta_L+\beta_M$, we have $-\underline{b}_M\leq\bar{b}_L$, and so $\delta_{-(t-1)}\leq\bar{b}_L$. But then, $\hat{\pi}_L(\underline{\delta},x_{-t},\delta_{-(t-1)})<0$ for all $\underline{\delta}$ by $(\hat{P}1)$. As with natural debate, when too few weak types of Moritz remain in period t, then Lones' marginal type strictly wants to concede. ‖

A.9. *Intransigence: Proof of Proposition 3*

STEP 1. *There exists a threshold $x(\eta)$ with $\lim_{\eta\downarrow 0} x(\eta) = -\infty$, such that debate cannot switch from ambivalence to intransigence when $x_t > x(\eta)$.*

Proof. Suppose otherwise, that for some finite x^* and arbitrarily small $\eta > 0$, there is a period t with $x_t > x^*$, where Moritz is ambivalent in t, but Lones is intransigent in $t+1$.

The decision payoff gain of Lones' cutoff type x_{t+1} when $m \in [x_t, x_{t+2}]$ concede to convict must then be balanced by the associated delay costs. Up to constant multiplicative factors, this gain and loss equal $\Delta(-\delta_{t+1}, \beta_L) f(x_{t+2}|x_{t+1})(x_{t+2} - x_t)$ and $k_L \eta (1 - F(x_{t+2}|x_{t+1}))$ respectively; thus, the step size $(x_{t+2} - x_t)$ is bounded above by a linear function in η.[50] But then, since x_t ambivalently wishes for x_{t+1} to prevail and Moritz' subsequent cutoff type x_{t+2} is only marginally stronger, his gross benefit from winning against Lones' types $\ell > x_{t+1}$ is of order η^2, as both the integrand and the domain of the positive area in Figure 2a are of order η. For vanishing η, the first-order delay costs outweigh x_{t+2}'s second order decision payoff gain. This contradiction refutes our assumption that debate switches from ambivalence in period t to intransigence in period $t+1$. ‖

STEP 2. *Natural ambivalent debate cannot last forever for small η.*

Proof. The indifference condition $\pi_i(\underline{\delta}, y, \bar{\delta}) = 0$ implies $\bar{\delta} > (1+\epsilon)\underline{\delta}$ for some $\epsilon > 0$ if $y - \underline{\delta} > 0$. For the bias β_i and falling density f raise the positive area in Figure 2a, while the delay costs—and thus the negative right hand tail—vanish. Thus, the propensity $\pi_i(\underline{\delta}, y, \bar{\delta})$ is positive for symmetric cutoff gaps, and indifference requires $\bar{\delta} > \underline{\delta}$; the uniform wedge $\epsilon > 0$ follows either from $\beta_i > 0$ or the bounded decay rate of the density f.[51]

But cutoff gaps $\delta_{t+1} \geq (1+\epsilon)\delta_t$ cannot grow exponentially in equilibrium (see Section 6).[52] ‖

STEP 3. *For any $\eta > 0$ and (σ, ∞)-equilibrium, there is a least t with $x_t > x(\eta)$ and $x_{t-1} > x(\eta)$. Let $t(\eta)$ the supremum such period t over all such equilibria. Then $\lim_{\eta\downarrow 0} \eta t(\eta) = 0$.*

Consider any $\eta > 0$ with $x(\eta) < 0$. In any odd period $s < t$, the benefit of holding out (the first term in (4)) is at most $(1+\beta_M)(F(x_{s+1}|x_s) - F(x_{s-1}|x_s))$, while the costs are at least $2k_M\eta(1 - F(x(\eta)|x_s))$, since $x_{s+1} < x(\eta)$. Thus, indifference implies

$$(1+\beta_M)(F(x_{s+1}|x_s) - F(x_{s-1}|x_s)) > 2k_M\eta(1 - F(x(\eta)|x_s)).$$

Given the log-submodular joint type distribution (cf. Section A.1) and $x_s < 0$, the distribution of Lones' types conditional on Moritz' (weak) type $m = x_s$ assigns relatively more weight to strong types of Lones $\ell \geq x(\eta)$, while the distribution conditional on Moritz' (stronger) type $m = 0$ assigns relatively more weight to Lones' weak conceding types, $\ell \in [x_{s-1}, x_{s+1}]$. So:

$$(1+\beta_M)(F(x_{s+1}|0) - F(x_{s-1}|0)) > 2k_M\eta(1 - F(x(\eta)|0)).$$

Adding over all odd periods $s < t(\eta) - 2$, we obtain

$$\frac{(1+\beta_M)(F(x(\eta)|0)}{k_M(1 - F(x(\eta)|0))} > (t(\eta) - 2)\eta.$$

As $\eta > 0$ vanishes, and $x(\eta)$ grows ever more negative, the LHS vanishes. ‖

All told, as η vanishes, $x_{t(\eta)} > x(\eta)$ in any (σ, ∞)-equilibrium with vanishing real time $t(\eta)\eta$, by Step 3; after period $t(\eta)$, debate is intransigent, as it can no longer switch from ambivalence to intransigence by Step 1, and cannot be ambivalent forever, by Step 2. This proves the first statement of Proposition 3. The second statement follows from part (*b*) of the next Lemma. ‖

For small period lengths $\eta > 0$, we can characterize cutoff gaps $\delta_t = x_t - x_{t-1}$, and thereby the sets of conceding types $[x_{t-1}, x_{t+1}]$, which measures the hazard rate of debate ending in period t.

50. Here, we use the fact that $x_t > x^*$ to bound below the hazard rate $f(x_{t+2}|x_{t+1})/(1 - F(x_{t+2}|x_{t+1}))$

51. Note how this argument fails for unbiased jurors, where communicative debate is ambivalent with *decreasing* cutoff gaps, as discussed after Theorem 5. The indifference condition then balances the positive effect of the decreasing density with the negative effect of delay costs. In contrast to the case with biased jurors, as delay costs get small, so do the cutoff gaps and hence the effect of the decreasing density.

52. This step bears some resemblance to the proof of Proposition 1. There, intransigence in period $t-1$ leads to the contradiction that the cutoff type x_t in the next period strictly prefers to concede. Here, ambivalence from period t onwards leads to the contradiction that the cutoff type $x_{t'}$ in some later period $t' > t$ strictly prefers to hold out. Since periods t and t' are possible far apart, this proof logic is more intricate. As a consequence, it does not yield an interpretable sufficient condition for intransigence, like condition (12) for ambivalence.

Lemma A.5. *Fix biases $\beta_L, \beta_M > 0$, type $x^* \in \mathbb{R}$, sufficiently small $\epsilon > 0$ and $\eta > 0$, any communicative equilibrium, and period t with $x_{t-2} > x^*$. Then (a) $\delta_t \geq -\bar{b}_{i(t)} + \epsilon$; (b) $x_{t+1} - x_{t-1}$ is of order η; (c) if $\beta_M = \beta_L \geq 0$ and $k_L = k_M > 0$, then $\delta_t > 0$, and (d) if $\beta_L \geq \beta_M$ and $k_L \leq k_M$ (with not both equal), then $\delta_t > 0$ for odd t and $\delta_t < 0$ for even t.*

Proof of Part (a). First, note that $\underline{b}_i, \bar{b}_i \to b_i$ as $\eta \downarrow 0$ in (10). By way of contradiction, assume that δ_t (for even t, say) is not bounded away from $-b_L$; this means that the *upper* corner of the staircase in Figure 5a is no more than ϵ above the *lower* boundary of the disagreement zone. Then Lones' maximal decision payoff gain $\Delta(-\delta_t, \beta_L)$ vanishes, while Moritz' gain $\Delta(-\delta_{t+1}, \beta_M)$ is boundedly positive, since $\delta_{t+1} = x_{t+1} - x_{t-1} + \delta_t > 0$. Thus, Lones' period t indifference condition, Moritz' period $t+1$ indifference condition, together with the fact that delay costs and the density functions in these indifference conditions differ by at most a constant factor, imply that the interval of Moritz' conceding types $[x_{t-1}, x_{t+1}]$ in period $t+1$ must exceed Lones' corresponding interval $[x_t, x_{t+2}]$ by the exploding factor $\Delta(-\delta_{t+1}, \beta_M)/\Delta(-\delta_t, \beta_L)$. But then $\delta_{t+2} = x_{t+2} - x_{t+1} = (x_{t+2} - x_t) - (x_{t+1} - x_{t-1}) + \delta_t$ is boundedly less than δ_t, that is, the upper corner of the staircase in Figure 5(a) is getting even closer to the lower boundary of the disagreement zone. Inductively, δ_{t+2s} eventually falls below $-b_L$, contradicting Lones' indifference condition in period $t+2s$ by property $(P1)$. This contradiction proves part (a). ‖

Proof of Part (b). By part (a), $\delta_t \geq -\bar{b}_{i(t)} + \epsilon$, and so $\Delta(\delta_t, \beta_{i(t)})$ is boundedly positive. But then the period t indifference condition implies that the interval length $x_{t+1} - x_{t-1}$ approximates $2\eta k_{i(t)}/(\Delta(\delta_t, \beta_{i(t)}) f(x_{t-1} | x_t, x \geq x_{t-1}))$ for small $\eta > 0$. ‖

Proof of Part (c). Assume to the contrary that $\delta_t \leq 0$, *i.e.* $x_t \leq x_{t-1}$, say for even t. Since $\delta_{t+1} > 0$, the same argument as in (a) implies that $[x_{t-1}, x_{t+1}]$ exceeds $[x_t, x_{t+2}]$ by a boundedly positive amount, so that $\delta_{t+2} < \delta_t \leq 0$, and δ_{t+2s} eventually falls below $-b_L$. ‖

Proof of Part (d). Assume, say, $\delta_t \geq 0$ for even t. By (a), δ_{t+1} is either negative or of order η, and so $\Delta(\delta_t, \beta_L)/k_L$ exceeds $\Delta(\delta_{t+1}, \beta_M)/k_M$ by a bounded amount. So the interval of Lones' conceding types $[x_t, x_{t+2}]$ exceeds Moritz' interval $[x_{t-1}, x_{t+1}]$ by a bounded amount. Then $\delta_{t+2} > \delta_t \geq 0$ and δ_{t+2s} eventually exceeds b_M, contradicting intransigence. ‖

Finally, we argue that the chance of Nixon-China debate $F(x_0)$ must vanish, too. Suppose otherwise, that x_0 is bounded below, for all η. Property $(\hat{P}3)$ in Section A.4 implies $\delta_0 = x_0 - x_{-1} > \bar{b}_M$. By $(\bar{P}3)$, Lones' initial indifference condition $\bar{\pi}_L(\delta_0, x_0, \cdot) = 0$ has up to two roots δ_1, an outer one above $\bar{b}_L$, and an inner one below. Steps 1 and 2 rule out the outer root in the proof of Proposition 3. The inner root is ruled out by an argument analogous to the one in step 1: The incentives of Lones' initial cutoff type x_0 to achieve an immediate conviction against types $m \in [x_{-1}, x_1]$ must be balanced by the incremental delay costs. So $x_1 - x_{-1}$ is of order η, but then Moritz' next cutoff type x_1 cannot be induced to hold out in period 1, as in Step 1. ‖

A.10. *Ambivalence: Proofs of Theorem 5 and Proposition 4*

Proof of Theorem 5. From the proof in Section A.5, recall the iterated domains $X(k)$ of cutoff pairs (x_0, x_1) consistent with drop-dead date k—or the communicative equilibrium for $k = \infty$—in the natural subgame. The proof sketch after the theorem statement provs that $X(k)$ has 'slope less than one', *i.e.* if $x_0 < x_0'$ and $(x_0, x_1), (x_0', x_1') \in X(k)$, then $x_1' - x_1 < x_0' - x_0$.

We argue that the corresponding iterated domain $\hat{X}(k)$ in the Nixon-China subgame has slope less than one too. The RHS of indifference condition (6) requires $\hat{\pi}_i(\delta_{-t}, x_{-t}, \delta_{-t+1}) = 0$. By $(\hat{P}3)$ in Section A.4, the indifference condition $\hat{\pi}_i(\cdot, x_{-t}, \delta_{-t+1}) = 0$ for given $y, \bar{\delta}$ may admit two roots $\underline{\delta}$— on either side of $-\underline{b}_i$. But given (12), only the outer ambivalence root $\underline{\delta} > -\underline{b}_i$ is compatible with equilibrium: For assume that Moritz' type x_{-t+1} is intransigent, $\delta_{-t+1} = x_{-t+1} - x_{-t} \leq -\underline{b}_M$. Then $\hat{\pi}_L(\underline{\delta}, x_{-t}, \delta_{-t+1}) < \pi_L(\delta_{-t+1}, -x_{-t}, \underline{\delta}) \leq \pi_L(\delta_{-t+1}, -x_{-t}, \bar{b}_L) < 0$, where the first inequality follows from $(\hat{P}6)$, the second from $(P3)$, and the last from $(P1)$, along with the fact that $\delta_{-t+1} \leq -\underline{b}_M \leq \bar{b}_M$, given $\pi_L(\bar{b}_M, -x_{-t}, \bar{b}_L) < 0$, by (12).

So motivated, let $\hat{\chi}_i(\bar{\delta}, y)$ be the unique root $\underline{\delta} > -\underline{b}_i$ of $\hat{\pi}_i = 0$, whenever one exists. The shooting function $\hat{\chi}_i$ decreases in y and increases in $\bar{\delta}$, with slope exceeding $1 + \varepsilon$ when y is small. Arguments analogous to those after Theorem 5 then imply that $\hat{X}(k)$ — the set of pairs (x_0, x_{-1}) consistent with a k-deferential equilibrium in the Nixon-China subgame—has 'slope less than one', *i.e.* if $x_0 < x_0'$ and $(x_0, x_{-1}), (x_0', x_{-1}') \in \hat{X}(k)$, then $x_{-1}' - x_{-1} < x_0' - x_0$.

For a contradiction, assume two (σ, τ)-equilibria with $x_0 < x_0'$. Then, $\delta_1 = x_1 - x_0 > x_1' - x_0' = \delta_1'$ and $\delta_0 = x_0 - x_{-1} < x_0' - x_{-1}' = \delta_0'$. Hence $0 = \bar{\pi}_L(\delta_0', x_0', \delta_1') > \bar{\pi}_L(\delta_0, x_0', \delta_1') > \bar{\pi}_L(\delta_0, x_0, \delta_1') > \bar{\pi}_L(\delta_0, x_0, \delta_1) = 0$, where the first inequality uses $(\bar{P}1)$ together with $\delta_0, \delta_0' > -\underline{b}_L$, the second uses $(\bar{P}2)$, and the third uses $(\bar{P}3)$ together with $\delta_1, \delta_1' > \bar{b}_L$. ‖

Proof of Proposition 4. We exploit the properties of the propensity function $\pi_i(\underline{\delta}, y, \bar{\delta})$. By $(P4)$, this rises in β_i and falls in κ_i, and by $(P3)$, it falls in $\bar{\delta}$ for ambivalent debate. Then the shooting function $\chi_i(\underline{\delta}, y)$ rises in β_i and falls in κ_i. Thus, for a fixed anchor x_0 and seed x_1, the entire cutoff sequence (x_t) rises in β_i and falls in κ_i. Finally, to restore the boundary

condition of the equilibrium cutoff vector (x_t)—namely, $x_\tau = \infty$ if τ is finite and $x_t \longrightarrow \infty$ if τ is infinite—x_1 falls if either juror grows more biased or more patient.

We finish the proof by analysing the Nixon-China subgame and the initial period indifference condition. Assume $\beta'_L \geq \beta_L$ and $\beta'_M \geq \beta_M$, with at least one strict. By the above, the iterated domain $X'(\tau)$ for parameters β'_i lies below the iterated domain $X(\tau)$ for parameters β_i, *i.e.* $x'_1 < x_1$ whenever $(\ell, x_1) \in X(\tau)$ and $(\ell, x'_1) \in X'(\tau)$ for some ℓ. Similarly, in the Nixon-China subgame, $\hat{X}'(\sigma)$ lies below $\hat{X}(\sigma)$, in that $x'_{-1} < x_{-1}$ whenever $(\ell, x_{-1}) \in X(-\sigma)$ and $(\ell, x'_{-1}) \in \hat{X}'(\sigma)$ for some ℓ, as seen in Figure 9b.

We first show that in equilibrium $x_0 > x'_0$. Assume not, so $x_0 \leq x'_0$. Then, $\delta'_1 = x'_1 - x'_0 < x_1 - x_0 = \delta_1$ since $X'(\tau)$ lies below $X(\tau)$ and both sets have "slope less than one". Similarly, $\delta'_0 = x'_0 - x'_{-1} > x_0 - x_{-1} = \delta_0$ since $\hat{X}'(\sigma)$ lies below the iterated domain $\hat{X}(\sigma)$, and both sets have "slope less than one". Writing Lones' initial propensity explicitly as an (increasing) function of his bias, we get the contradiction $0 = \bar{\pi}_L(\delta'_0, x'_0, \delta'_1; \beta'_L) > \bar{\pi}_L(\delta_0, x'_0, \delta'_1; \beta'_L) \geq \bar{\pi}_L(\delta_0, x_0, \delta'_1; \beta'_L) > \bar{\pi}_L(\delta_0, x_0, \delta_1; \beta'_L) \geq \bar{\pi}_L(\delta_0, x_0, \delta_1; \beta_L) = 0$.

We next show that $x_t > x'_t$ for all $t \in [1, \tau - 1]$. If not, there is an earliest time $T < \tau$ with $x_T \leq x'_T$. Then $\delta_T < \delta'_T$, and inductively $x_s < x'_s$ and $\delta_s < \delta'_s$ for all $s > T$ since the shooting function $\chi_i(\underline{\delta}, y)$ increases in both arguments, and in the bias, contradicting the assumption that $x_\tau = x'_\tau = \infty$. Similarly, a drop-dead date σ in the Nixon-China subgame requires $x_t > x'_t$ for all $t \in [-\sigma + 1, -1]$. This completes the proof that natural debate slows down but Nixon-China debate speeds up as either juror grows more biased.

For unbiased jurors, both natural and Nixon-China debate slow down as either juror grows more patient. For without bias, the two subgames are symmetrical, and so uniqueness requires $x_0 = 0$. The proof sketch after the proposition is then a proof. ‖

A.11. *Asymptotic Stationarity: Proof of Theorem 6*

Given the asymptotic type density $f^\infty(\delta | \delta \geq -\underline{\delta})$ in Lemma A.2, the asymptotic propensity is:

$$\pi_i^\infty(\underline{\delta}, \bar{\delta}) = \int_{-\underline{\delta}}^{\bar{\delta}} (\Delta(\delta, \beta_i) - \kappa_i) f^\infty(\delta | \delta \geq -\underline{\delta}) d\delta - \int_{\bar{\delta}}^{\infty} 2\kappa_i f^\infty(\delta | \delta \geq -\underline{\delta}) d\delta. \tag{31}$$

by Section 3.5. After the theorem, we proved that the asymptotic indifference curves $\pi_M^\infty(\delta_{MJ}, \delta_{EI}) = 0$ and $\pi_L^\infty(\delta_{EI}, \delta_{MJ}) = 0$ cross at most once. Indeed they cross: Moritz' indifference $\pi_M^\infty(\delta_{MJ}, \delta_{EI}) = 0$ admits a solution δ_{MJ} for every δ_{EI}, but for all δ_{EI} we have $\pi_M^\infty(\delta_{MJ}, \delta_{EI}) < 0$ for all δ_{EI} if $\delta_{MJ} < -\underline{b}_M$ (then the positive area G in Figure 2a does not exist), and $\pi_M^\infty(\delta_{MJ}, \delta_{EI}) > 0$ if δ_{MJ} is sufficiently large (since then the positive area G in Figure 2a outweighs the negative area L as the density diminishes exponentially). Thus, Moritz' indifference curve connects the left and right boundary in Figure 7a. Similarly, Lones' indifference curve $\pi_L^\infty(\delta_{EI}, \delta_{MJ}) = 0$ connects the upper and lower boundary in Figure 7a. Hence, the indifference curves cross.

Next, we fix a communicative equilibrium, and argue that the cutoff gaps must converge to the limit gaps, *i.e.* $\delta_{MJ} \equiv \lim_{t \to \infty} \delta_{2t+1}$ and $\delta_{EI} \equiv \lim_{t \to \infty} \delta_{2t}$. For large t, equilibrium cutoff gaps are bounded above and below, so it suffices to show that the (closed) set $\mathcal{D}_{MJ}$ of accumulation points of $\{\delta_{2t+1}\}$ is degenerate, equal to $\{\delta_{MJ}\}$, and similarly $\mathcal{D}_{EI} = \{\delta_{EI}\}$. Assume otherwise, that $\max \mathcal{D}_{MJ} > \min \mathcal{D}_{MJ}$. Then, there exist $\delta, \delta' \in \mathcal{D}_{EI}$ with $\pi_M^\infty(\min \mathcal{D}_{MJ}, \delta) = 0$ and $\pi_M^\infty(\max \mathcal{D}_{MJ}, \delta') = 0$, and (*P*5) implies $(1+\varepsilon)|\max \mathcal{D}_{MJ} - \min \mathcal{D}_{MJ}| < |\delta' - \delta|$ which in turn is bounded above by $|\max \mathcal{D}_{EI} - \min \mathcal{D}_{EI}|$. But the symmetric argument with Lones' indifference curve then leads to the contradiction that $(1+\varepsilon)|\max \mathcal{D}_{EI} - \min \mathcal{D}_{EI}| < |\max \mathcal{D}_{MJ} - \min \mathcal{D}_{MJ}|$. The arguments for the Nixon-China subgame are analogous. ‖

A.12. *Devil's Advocates: Proof of Proposition 8*

First, as in (12), small biases β_L and β_M imply

$$P_M(y - \bar{b}_L, y) < 0 \qquad \text{and} \qquad P_L(y - \bar{b}_M, y) < 0 \qquad \text{for all } y \in \mathbb{R}. \tag{32}$$

These inequalities hold when $\beta_L = \beta_M = 0$ and extend to small biases by continuity.[53]

Next, by Lemma A.5(b), the interval length of conceding types $x_{t+2} - x_t$ (for x_t above some fixed threshold x^*) vanishes for small $\eta > 0$, and $\delta_{2t} + \bar{b}_L \geq \epsilon$ and $\delta_{2t+1} + \bar{b}_M \geq \epsilon$ for some $\epsilon > 0$. Graphically, the cutoff staircase of Figure 5a is inside the disagreement zone $\{(\ell, m) : m - \bar{b}_L \leq \ell \leq m + \bar{b}_M\}$, and does not approach its boundaries.

Let $\eta > 0$ be small enough that the interval length $x_{t+2} - x_t < \epsilon/(T^* + 1)$. Defining t by $x_{2t-1} \leq \ell - \bar{b}_M < x_{2t+1}$, dictator ℓ is ready to acquit in period $2(t+1)$, or possibly earlier, by (32). But debator ℓ plays devil's advocate and wishes to convict after period $2(t+T^*)$, since $\ell \geq x_{2t-1} + \bar{b}_M \geq x_{2t+1} - \delta_{2t+1} + \epsilon - (x_{2t+1} - x_{2t-1}) = x_{2t} + \epsilon - (x_{2t+1} - x_{2t-1}) \geq x_{2(t+T^*)}$. ‖

53. For the bias upper bound implied by (32) is tighter for higher hazard rates of the type distribution.

Acknowledgments. We thank the editor and three anonymous referees for their guidance, and Andy Atkeson, Simon Board, Ra'anan Boustan, Ed Green, Faruk Gul, PJ Lamberson, Maximo Langer, Markus Mobius, Mike Peters, Ariel Rubinstein, Alexander Stremitzer, Bruno Strulovici, and especially William Zame for helpful comments. We have also benefited from comments at Bonn, CalTech, Chicago-Booth, Duke-UNC, ESEM 2008, ESSET 2012 and 2015, Games 2008, Microsoft Research, Penn State, SED 2008, SITE 2014, SWET 2008, UBC, and UC Davis. Menghan Xu provided excellent research assistance. The usual disclaimer applies.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

ABREU, D. and GUL, F. (2000), "Bargaining and Reputation", *Econometrica*, **68**, 85–117.
AUMANN, R. and HART, S. (2003), "Long Cheap Talk", *Econometrica*, **71**, 1619–1660.
AUSTEN-SMITH, D. and BANKS, J. (1996), "Information Aggregation, Rationality, and the Condorcet Jury Theorem", *American Political Science Review*, **90**, 34–45.
AUSTEN-SMITH, D. and FEDDERSEN, T. (2006), "Deliberation, Preference Uncertainty, and Voting Rules", *American Political Science Review*, **100**, 209–217.
CHAN, J., LIZZERI, A., SUEN, W., *et al*. (forthcoming), "Deliberating Collective Decisions", *Review of Economic Studies*.
CHE, Y.-K. and KARTIK, N. (2009), "Opinions as Incentives", *Journal of Political Economy*, **117**, 815–860.
CHO, I.-K. (1987), "A Refinement of Sequential Equilibrium", *Econometrica*, **55**, 1367–1389.
COUGHLAN, P. J. (2000), "In Defense of Unanimous Jury Verdicts: Mistrials, Communication, and Strategic Voting", *American Political Science Review*, **94**, 375–393.
CRAWFORD, V. P. and SOBEL, J. (1982), "Strategic Information Transmission", *Econometrica*, **50**, 1431–1451.
DAMIANO, E., LI, H. and SUEN, W. (2012), "Optimal Deadlines for Agreements", *Theoretical Economics*, **7**, 357–393.
DEWATRIPONT, M. and TIROLE, J. (1999), "Advocates", *Journal of Political Economy*, **107**, 1–39.
ED CARNES, C. C. J. (2016), "Eleventh Circuit Pattern Jury Instructions". http://www.ca11.uscourts.gov/pattern-jury-instructions.
FEDDERSEN, T. and PESENDORFER, W. (1996), "The Swing Voter's Curse", *American Economic Review*, **86**, 408–424.
—— (1997), "Voting Behavior and Information Aggregation in Elections with Private Information", *Econometrica*, **65**, 1029–1058.
FORGES, F. (1990), "Equilibria with Communication in a Job Market Example", *Quarterly Journal of Economics*, **105**, 375–398.
FUDENBERG, D. and TIROLE, J. (1991), *Game Theory* 1st edn. (Cambridge, MA: MIT Press).
GERARDI, D. and YARIV, L. (2007), "Deliberative Voting", *Journal of Economic Theory*, 134, 317–338.
—— (2008), "Information Acquisition in Committees", *Games and Economic Behavior*, 62, 436–459.
GERSHKOV, A. and SZENTES, B. (2009), "Optimal Voting Schemes with Costly Information Acquisition", *Journal of Economic Theory*, **144**, 36–68.
GOLTSMAN, M., HORNER, J., PAVLOV, G., *et al*. (2009), "Mediation, Arbitration and Negotiation", *Journal of Economic Theory*, **144**, 1397–1420.
GROSS, S. R., POSSLEY, M. and STEPHENS, K. (2017), "Race and Wrongful Convictions in the United States" (National Registry of Exonerations).
GUL, F. and LUNDHOLM, R. (1995), "Endogenous Timing and the Clustering of Agents Decisions", *Journal of Political Economy*, **103**, 1039–1066.
GUL, F. and PESENDORFER, W. (2012), "The War of Information", *Review of Economic Studies*, **79**, 707–734.
IARYCZOWER, M., SHI, X. and SHUM, M. (forthcoming), "Can Words Get in the Way? The Effect of Deliberation in Collective Decision-Making", *Journal of Political Economy*.
KARLIN, S. and RINOTT, Y. (1980), "Classes of Orderings of Measures and Related Correlation Inequalities. I. Multivariate Totally Positive Distributions", *Journal of Multivariate Analysis*, **10**, 467–498.
KRISHNA, V. and MORGAN, J. (2004), "The Art of Conversation: Eliciting Information from Experts through Multi-stage Communication", *Journal of Economic Theory*, **117**, 147–179.
LI, H., ROSEN, S. and SUEN, W. (2001), "Conflicts and Common Interests in Committees", *American Economic Review*, **91**, 1478–1497.
LI, H. and SUEN, W. (2009), "Decision-making in Committees", *Canadian Journal of Economics*, **42**, 359–392.
MEYER-TER-VEHN, M., SMITH, L. and BOGNAR, K. (2017), "Like Minded Debate" (Mimeo).
PERSICO, N. (2003), "Committee Design with Endogenous Information", *Review of Economic Studies*, **70**, 1–27.
PIKETTY, T. (2000), "Voting as Communicating", *Review of Economic Studies*, **67**, 169–191.
RILEY, J. (1980), "Evolutionary Equilibrium and The War of Attrition", *Journal of Theoretical Biology*, **82**, 383–400.
SMITH, L., SORENSEN, P. and TIAN, J. (2016), "Informational Herding, Optimal Experimentation, and Contrarianism" (Mimeo).