# *A Conversational War of Attrition*[*]

Katalin Bognar[†] Private Sector

Moritz Meyer-ter-Vehn[‡] UCLA

Lones Smith[§] Wisconsin

September 30, 2013

**Abstract**

We explore costly dynamic deliberation by two like-minded but differentially informed individuals who must agree on a common verdict. We offer two seemingly opposed findings. As the conversation transpires and increasingly strong types of the players concede to their opponent's verdict it becomes more and more likely that the conversation is moot ex post. Yet, ex ante, among all equilibria of the game, the one with the longest possible conversation is most efficient and uniquely obeys forward induction. As communication costs $\kappa$ vanish, information is aggregated perfectly in this equilibrium; while the conversation persists for an exploding number of periods, total waiting costs also vanish at rate $\kappa^{2/3}$.

[*]We thank Simon Board, Ed Green, Markus Mobius, Mike Peters, and Bruno Strulovici for helpful comments. We have also benefited from comments at SWET 08, Penn State, UC Davis, SED 08, Games 08, ESEM 08, CalTech, UBC, Duke-UNC, ESSET 2012, Microsoft Research. The usual disclaimer applies. Keywords: Committee decision making, Cheap talk, Deliberation, Juries, Monotone methods, War of attrition. JEL codes: D71, D72, D82, D83, C72

[†]katalin.bognar@gmail.com

[‡]mtv@econ.ucla.edu; Moritz thanks the NSF for financial support of this research.

[§]lones@ssc.wisc.edu; Lones thanks the NSF for financial support of this research.

# 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*

Economics is not in the business of disputing tastes. And few topics pique an economist so much as seeing how preference diversity is either resolved, or proven unresolvable. But even like-minded individuals disagree if they are differentially informed. We explore the dynamics of a costly deliberation by two such people who cannot leave a room until they agree. We have in mind juries, tenure cases, or panels, whose members share an imprimatur to dispassionately arrive at the truth. Of course, if they could simply share information, they would be done with it. We should then not observe lengthy disagreements. We instead study a model with gradual information revelation.

We have in mind a fully Bayesian model of communication followed by a binary choice, with the smart action dependent on a binary true state, like guilty or innocent. The communication is either hamstrung by a coarse language, or individuals are cognitively unable to adduce their precise signal. Since costly talk eventually must "get real", we simply assume that it is *dispositive* from the outset. In the resulting world, there is a sequence of yays or nays until a concurring irreversible decision. Still, this framework can be seen as ($a$) extending the pivot voting literature, allowing for dynamic learning from disagreement; ($b$) embellishing the dynamic cheap talk literature with deliberation costs; or ($c$) adding dynamic aggregation of private information to the more recent literature on committee voting.

What emerges in our setting is either immediate agreement, or an incomplete information war of attrition where players "doth protest" disagreement until one concedes to the other's position. We focus on the latter subgame, with juror Moritz pushing for acquittal and juror Lones for conviction. If we express each juror's type as an absolute log-likelihood ratio of guilt over innocence, then the type difference is the de facto true state — in other words, only it can be learned. The cost of communication conceals such Solomonic wisdom. An equilibrium is instead a sequence of thresholds, one per period (Theorem 1) for each player. A juror concedes once his threshold surpasses his signal; otherwise he pushes on. An equilibrium can be portrayed as a zig-zag path through the type space. As public finance economists are familiar with, triangular deviations from the diagonal capture the magnitude of decision errors.

Apart from a class of contrarian conversations that we discuss in the appendix, we characterize all sequential equilibrium conversations, indexing them by their "drop-dead" dates (Theorem 3). In a *deferential equilibrium*, one player defers to the other at some stage; this follows if his rival "digs in his heels". These equilibria correspond to arguments that terminate either by a protocol or cloture rule — such as occurs in parliamentary debates or "fiscal cliff" talks in the USA. Exploiting our common interests structure, we argue more strongly that deferential equilibria are efficient *conditional* on the horizon length (Theorem 2). But unlike the analogous asymmetric outcomes of a war of attrition, these are unconstrained inefficient outcomes. The focal equilibrium in our paper has no certain last period. It represents the most refined communication. We show that this *communicative equilibrium* is unconstrained Pareto optimal. We also argue that it is the only stable equilibrium (Theorem 4). The basic logic is that ours is a game of common interests, and thus the Pareto test roughly coincides with the selfish stability test imposed by forward induction. Intuitively, digging in one's heels is not forwardly rational, since types who deviate in the next period can convey their extremely powerful private signals.

Our model has a remarkable amount of freedom in constructing equilibria. An equilibrium is a potentially infinite sequence of thresholds satisfying a nonlinear second order difference equation. It may therefore come as a surprise that we can argue uniqueness. For when the players' log-likelihood ratio has a log-concave density, we show that deferential and communicative equilibria are each unique for each horizon (Theorem 5).[1] Our proof for the communicative equilibrium argues that a unique starting point for the conversation obeys the transversality condition.

Our model captures two commonly claimed features of dynamic communication. First, a third party observing the debate would see his posterior belief about the smart action course zig zag each period, always agreeing with the last proposal. Second, dynamic communication embeds an option value of changing one's mind. With low communication costs, individuals in equilibrium argue for one course of action, even though they may believe in the wisdom of the opposite course of action. This divergence between how the jurors vote as debater, and how they would act as a dictator, offers a simple rationalization for the oft-claimed "devil's advocate" play by some individuals.

We next explore a dissonance between ex ante and ex post efficiency. We prove that as the conversation transpires, it grows increasingly likely that it will is "moot" ex post (Proposition 1). In the end, the longest conversations will prove the most in-

[1]This condition, adapted from Smith, Sorensen, and Tian (2012), is met by the standard continuous signal distributions employed by economists.

consequential. But we then find that arguing for as long as possible is the most efficient outcome (Proposition 3). We argue that the resulting ex post negative correlation between investment expenditures and total returns to arguing is not only a feature of our most efficient equilibrium, but also of dynamic experimentation. The logic is that the lengthiest conversations are really the failure events of the dynamic option of learning (Weitzman 1979). This finding should prove a cautionary tale to those exploring the returns to pure research.

We finally revisit our communication costs. The meaning of the yay or nay votes obviously depends on the deliberation cost. A smaller cost permits more refined communication. As the per period delay cost $\kappa$ shrinks, so do the welfare loss triangles, and information aggregation consequently perfects. We first argue that in the communicative equilibrium, the conversation grows more refined in an "accordion sense" — namely, all signal thresholds compress, and separation of types is more refined (Proposition 4). We then turn to the limit where the cost $\kappa$ vanishes. Here we argue that the expected number of periods of conversation explodes, but at a rate less than inverse to the cost $\kappa$. Since marginal costs are proportional to the areas of the welfare loss triangles, we argue that the total waiting costs vanish at rate $\kappa^{2/3}$ (Proposition 5).

Related Literature. Our paper relates to and builds on several literature strains. A primary one is the study of strategic voting, as pioneered by Austen-Smith and Banks (1996) and Feddersen and Pesendorfer (1996), (1997), (1998). These papers observe that sincere voting may not be an equilibrium, even among like-minded but differentially informed voters; rather, rational individuals condition their vote on the event that their vote is pivotal. Departing from these papers, we explore the role of costly pre-vote communication that allows jurors to signal the strength of their information through delay. The pivotal event on which jurors then condition their decisions in every period is that their colleague is about to concede. Our paper might be seen as a critique of conclusions based on these static models, since they pool together strong and very strong signals for or against a position, while we allow them to separate by holding out.

There is also a large literature on cheap talk, pioneered by Crawford and Sobel (1982). This has led into a study of dynamic communication. For a salient example tangentially related to our thrust, Aumann and Hart (2003) study cheap talk that transpires over very long horizons. Their conversation is used for coordinating play in a one-shot game, but is entirely non-dispositive and perfectly free.

The committee decision literature, surveyed in Li and Wing (2009), addresses the role of free pre-vote communication. Coughlan (2000) observes that with common pref-

erences and binary information, a straw vote suffices to elicit private information and avoids strategic behavior by the jurors.[2] Austen-Smith and Feddersen (2006) counter that jurors with different preferences would not act truthfully in the straw vote; Li, Rosen, and Wing (2001) and Gerardi and Yariv (2007) extend this argument by studying the maximal amount of information elicited by a straw vote. Our paper maintains Coughlan (2000)'s assumption of common preferences, and extends his analysis by introducing coarse costly communication. Unlike ours, these papers are formally static.

A different strand of the committee literature focuses on information acquisition. 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. Lizzeri and Yariv (2012) model joint information acquisition by a committee, in the spirit of Wald (1947), and focus on the *deliberation rule* that determines when information acquisition stops and the vote is taken. By contrast, our jurors do not acquire information from outside sources but learn about each other's type.

There is also a recent literature on search by committee — Albrecht, Anderson, and Vroman (2010), Compte and Jehiel (2010), and Moldovanu and Shi (2013) — where committees sequentially sample alternatives and vote each period on whether to stop or continue the search. In contrast to these papers, our focus is on dynamic learning about a fixed state of the world. Strulovici (2010) studies experimentation by committee in a bandit model, where each player is concerned that experimentation may reveal information that others will use against him. This is not a concern in our model where jurors learn each other's type.

Our game resembles a standard war of attrition with incomplete information. In other wars of attrition, e.g., Gul and Pesendorfer (2012), to stop is to lose the prize. By contrast, in our model the act of stopping signifies that a juror almost agrees with his peer. In the end, the conflict nearly disappears. As a result, costly delay is optimal in our model, and cloture rules that terminate conversations at a fixed date only serve to lower welfare. In contrast, deadlines generally increase efficiency when there is a conflict of interest, as in Gul and Lundholm (1995) and Damiano, Li, and Suen (2012). Our jurors, by contrast, have no conflict of interest, and only disagree only because they possess divergent information.

We prove most results in the appendix.

[2]Piketty (2000) explores the communication role of repeated voting in multi-stage elections.

## 2 The Model

The Extensive Form Game. Two jurors Lones and Moritz alternate arguing in periods $t = 0, 1, 2, \ldots$ to convict or acquit, $\mathcal{C}$ or $\mathcal{A}$, the defendant of a crime. Moritz proposes a verdict in period zero. Lones replies in period one with his own proposal. If he agrees, the game ends; otherwise, Moritz responds in period two with a proposed verdict, and so on. The game ends with any consecutive concurring verdicts. We can interpret this game form as a sequential voting game with a unanimity rule.

The defendant is either *guilty* or *innocent*. Jurors share a flat common prior on the state of the world $\theta = \mathcal{G}, \mathcal{I}$. Before the game, each juror privately observes a signal about the defendant's guilt. These signals $\lambda, \mu$ are private beliefs in $[0, 1]$ about the guilty state, and conditionally iid. We assume no perfectly revealing signal, and a strictly positive unconditional density on $(0, 1)$ that is symmetric about $1/2$.

Jurors share the same cardinal preferences over *outcomes*: They want to take *the right decision*, i.e. convict the guilty and acquit the innocent. They also share the same costs for each period of delay. *Decision costs* are 0 for the right decision, and 1 for the wrong one. *Waiting costs* are equal to 1 per unit time and the length of a period is $\kappa < 1$. The jurors are risk-neutral and do not discount future payoffs — they try to minimize the expected sum of waiting costs and decision costs.

So the game resembles a war of attrition: a stopping game in which a player trades off the exogenous cost of continuing against the strategic incentives to reach his preferred verdict. But this preferred verdict may change as he learns his peer's type.

Reformulating Signals. It will simplify matters to represent the signals $\lambda, \mu$ in a non-standard fashion, as log-likelihood-ratios, and with a different reference state for each juror. So the transformed jurors' *types* are $\ell = \log(\lambda/(1-\lambda)), m = \log((1-\mu)/\mu)$. Let $\Gamma(\mathcal{G}|\ell, m)$ be the conditional probability of guilt, derived from a joint distribution on the larger space $\Omega = \{\mathcal{G}, \mathcal{I}\} \times \mathbb{R}^2$ in which $(\theta, \ell, m)$ lives. Bayes' rule implies:

$$\pi(\ell, m) = \Gamma(\mathcal{G}|\ell, m) = \frac{e^{\ell - m}}{e^{\ell - m} + 1} \tag{1}$$

Conviction is the *ex-post correct* verdict iff $\pi(\ell, m) \geq \frac{1}{2}$. As $\pi$ depends on the types $(\ell, m)$ only via the *type differential* $\delta \equiv \ell - m$, this inequality holds when $\delta \geq 0$.

Let $f(x)$ denote the unconditional signal type density of $x = \ell, m$. Then $f > 0$ and $f$ is symmetric about 0 by our assumptions on $\lambda$ and $\mu$. We also assume that $f$ is log-concave.[3]

[3]For instance, if the signals $\lambda, \mu$ are uniform on $[0, 1]$, then $f(x) = e^x/(1 + e^x)^2$ is log-concave.

The incentives of a juror with signal $y$ depend on the conditional density $f(x|y)$ over his colleague's type $x$, or equivalently, on the conditional density $f(y+\delta|y)$ over the type differential $\delta = x - y$. Let $h(x,y)$ be the joint density and define the *correlation* $r(x,y) \equiv h(x,y)/(f(x)f(y))$. Then Bayes' rule implies $f(x|y) = f(x)r(x,y)$. We show in the appendix 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 of the state are affiliated,[4] so too are their log-likelihood ratios; but then the inversely defined types $\ell, m$ are negatively affiliated. We now assert a log-submodularity relation in the same spirit.

**Lemma 1** *The conditional density $f(y+\delta|y)$ is log-submodular in $(y,\delta)$.*

Strategies and Payoffs. After Moritz proposes his initial verdict, the game ends as soon as a juror agrees with the previous verdict. We thus describe each player's pure strategies by the planned *stopping times* — the first period he will agree with his predecessor.[5] The game has two subgames $\mathcal{A}$ and $\mathcal{C}$ — labeled by Moritz's initial vote.[6] So Lones has a strategy described by two (odd) periods in which he first concedes to Moritz after each initial proposal, while Moritz' strategy consists of his initial proposal and his planned (even) concession period.

Call two strategy profiles *equivalent* if they almost surely prescribe the same outcome of the game; for a given strategy of one juror call two strategies of the other juror *equivalent* if the resulting strategy profiles are equivalent. We solve for the Bayes Nash equilibria (BNE); we later prove that any BNE is equivalent to a sequential equilibrium.

Consider either player, say with type $y$, faced with a rival of type $x$ with strategy $\varsigma(x)$. His expected costs of stopping in period $t$ are:

$$\sum_{s=1}^{t-1} \int_{\{x|\varsigma(x)=s\}} [1-\pi(y,x)+s\kappa] f(x|y)dx + \int_{\{x|\varsigma(x)>t\}} [\pi(y,x)+t\kappa] f(x|y)dx \qquad (2)$$

These terms respectively reflect the decision and waiting costs when type $y$ eventually prevails or concedes. In the first case, he prevails in some period $s < t$, and the decision costs are the chance that his colleague was correct $1-\pi$. In the second case, he concedes in period $t$, and the decision costs switch to $\pi$.

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

[5]As usual, we assume that a player who stops in period $t$ also plans to stop at all later periods $t' > t$ that are ruled out by his own earlier actions. Similarly, we assume that Moritz plans to stop at any period $t$ after the initial verdict which he does not choose.

[6]Of course, since there is an initial move by Nature, this extensive form has no proper subgames. We will however use this term for what Kreps and Wilson (1982) refer to as the *subform*.

# 3 Equilibrium Analysis

## 3.1 Monotonicity of Best Responses

The initial vote by Moritz fixes an endogenous ordering on types. In the $\mathcal{A}$-subgame, let us call higher types of Lones and Moritz *stronger*. Indeed, a stronger type of a juror is more convinced that he is arguing for the right verdict. In the $\mathcal{C}$-subgame, *stronger* types of Lones and Moritz are both lower. Call a strategy in the stopping subgame *monotone* if whenever some type of a juror holds out until period $t$, a stronger type holds out until period $t$ with certainty.

**Lemma 2 (Monotonicity in Subgames)** *A* single crossing property *holds: if any type of a juror prefers to hold out from $t$ to $t' > t$, then any stronger type prefers to do so, and strictly so if period $t$ is reached with positive probability. So every best response strategy of a juror to any strategy of his colleague is equivalent to a monotone strategy.*

One line of thinking about Lemma 2 argues that higher types are not only more convinced of their guess about the state, but also more convinced that their opponent does not entertain as strong an opposing signal — due to the negative correlation. The proof takes a different road: By conditioning on the state $\theta$, the proof shows that the benefit of holding out from period $t$ to $t'>t$ satisfies single-crossing.

*Proof of Lemma 2:* Consider Moritz' choice to hold out until period $t$ or $t'$ in the $\mathcal{A}$-subgame. If Lones concedes in period $s \in \{t+1, \cdots, t'-1\}$, then holding out until $t'$ increases decision costs by 1 if the state is $\mathcal{G}$, and reduces decision costs by 1 if the state is $\mathcal{I}$. Also, holding out increases waiting costs by $(s-t)\kappa$. 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 $(t'-t)\kappa$. Thus, when Lones' stopping time is $\varsigma(\ell)$, holding out until period $t'$ increases the expected costs of Moritz's type $m$ by:

$$\begin{array}{l} G(\mathcal{G}|m)\left[\sum_{s=t+1}^{t'-1}\int_{\{\ell:\varsigma(\ell)=s\}}(1+(s-t)\kappa)f^{\mathcal{G}}(\ell)d\ell+\int_{\{\ell:\varsigma(\ell)>t'\}}(t'-t)\kappa f^{\mathcal{G}}(\ell)d\ell\right] \\ +G(\mathcal{I}|m)\left[\sum_{s=t+1}^{t'-1}\int_{\{\ell:\varsigma(\ell)=s\}}(-1+(s-t)\kappa)f^{\mathcal{I}}(\ell)d\ell+\int_{\{\ell:\varsigma(\ell)>t'\}}(t'-t)\kappa f^{\mathcal{I}}(\ell)d\ell\right] \end{array} \quad (3)$$

The first line is positive — for in state $\mathcal{G}$, if Moritz argues longer for $\mathcal{A}$, both 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 (3) are negative, then the second line must be negative. When $m$ increases, costs (3) remain negative since $G_m(\mathcal{I}|m) = -G_m(\mathcal{G}|m) > 0$. So we have established a *single crossing property*: if $m$ prefers to hold out from $t$ to $t'$, then so does any type $m' > m$.

Thus, Moritz' best response in the $\mathcal{A}$-subgame increases in his type; similar arguments apply to the $\mathcal{C}$-subgame, and to Lones' strategies after $\mathcal{A}$ and $\mathcal{C}$. $\square$

Lemma 2 yields a standard skimming-property of equilibria: More extreme types quit in every period until communication ends. Lones' monotone strategy in the $\mathcal{A}$-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 the supremum type of Lones that concedes by period $t$. Moritz' monotone strategy in the $\mathcal{A}$-subgame is described by a weakly increasing sequence of even-indexed cutoff types $(x_t)_{t\in 2\mathbb{N}+2}$; monotone strategies in the $\mathcal{C}$-subgame are described by weakly decreasing sequences $(x_{-t})_{t\in 2\mathbb{N}+1}$ and $(x_{-t})_{t\in 2\mathbb{N}+2}$. *Hereafter, we assume monotone strategies, with cutoff types holding out.*

We now step back and consider Moritz' behavior in period zero. Call a strategy of Moritz *contrarian* if there exists a pair of types, with the lower type voting for acquittal and the higher type for conviction; otherwise, Moritz' strategy is *sincere*. Also, call a strategy of Moritz *responsive* if not all types vote for the same verdict. So a strategy of Moritz is sincere and responsive exactly when there exists a finite cutoff type $x_0$, with higher types initially proposing $\mathcal{A}$, and lower types initially proposing $\mathcal{C}$. Then, by definitions of our cutoffs, $x_{-2} \leq x_0 \leq x_2$.

Call a strategy of Lones *contradictory* if a positive measure of types disagrees with Moritz in period one no matter what Moritz proposed in period zero. Otherwise, if almost all types agree with some proposal of Moritz, Lones is *agreeable*. Formally, a monotone strategy of Lones is agreeable exactly when his cutoffs obey $x_{-1} \leq x_1$.[7]

**Lemma 3 (Sincerity and Agreeability)** *Any best response of Moritz to an agreeable, monotone strategy of Lones is sincere. Conversely, any best response of Lones to a sincere, monotone strategy of Moritz is equivalent to an agreeable strategy. Thus, up to equivalence, in equilibrium Moritz is sincere if and only if Lones is agreeable.*

The paper focuses on sincere agreeable responsive equilibria, and will touch on the insights from contrarian, contradictory equilibria only in §A. By monotonicity in Lemma 2, any sincere agreeable responsive equilibrium is characterized (as seen in Figure 1) by a cutoff sequence $(x_t)_{t\in\mathbb{Z}}$ with $|x_0| < \infty$, and:

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

[7]For then almost all types concede in period one to some proposal of Moritz. When $x_{-1} > x_1$, then types $[x_1, x_{-1}]$ always contradict Moritz in period one.

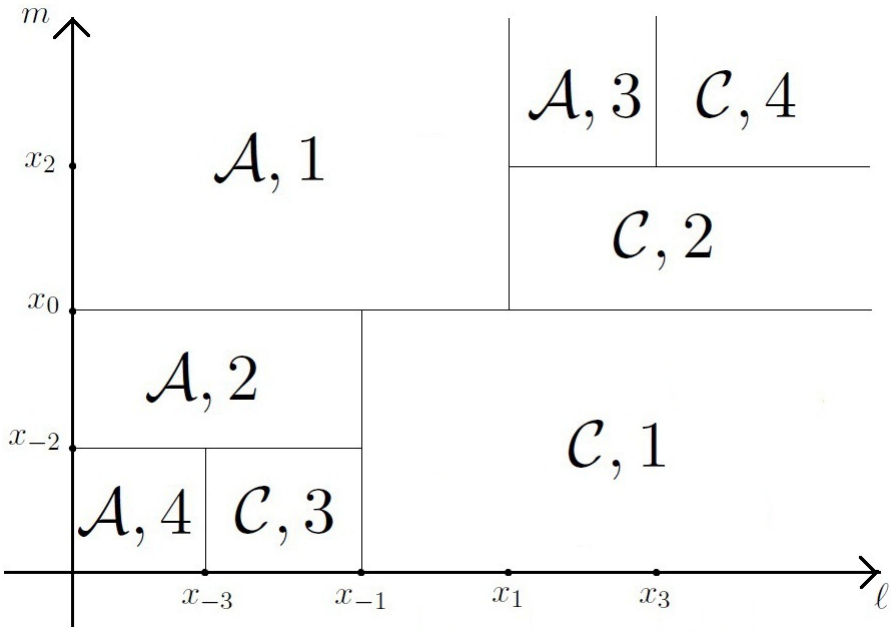


Figure 1: **Sincere and Agreeable Strategies.** This illustrates the outcome of a sincere agreeable strategy profile. Lones' types and (odd) cutoffs are on the horizontal axis, and Moritz' types and (even) cutoffs are on the vertical axis.

## 3.2 Equilibrium Characterization and the Zick-Zack Property

We now characterize equilibrium cutoffs $(x_t)$ in terms of indifference conditions. Define the *propensity to hold out* in the $\mathcal{A}$-subgame as a juror's expected payoff gain from holding out one more period — for instance, for juror type $y$ when his peer juror's type is above $x_{t-1}$, and types $x \in [x_{t-1}, x_{t+1}]$ will next concede (as seen in Figure 2):

$$P^{\mathcal{A}}(x_{t-1}, y, x_{t+1}) = \int_{x_{t-1}}^{x_{t+1}} (2\pi(y,x) - 1 - \kappa) f(x|y)dx - \int_{x_{t+1}}^{\infty} 2\kappa f(x|y)dx \qquad (5)$$

Holding out increases waiting costs by $\kappa$ if one's colleague concedes in the subsequent period, and by $2\kappa$ otherwise; the benefit owes to a reduction of decision costs when $2\pi(y,x) - 1 = \frac{e^{y-x}-1}{e^{y-x}+1} > 0$, namely, when $y > x$. Ex-post, this cost reduction exceeds the waiting cost $\kappa$ if $y - x \geq k \equiv \log \frac{1+\kappa}{1-\kappa}$.

Similarly, define $y$'s propensity to hold out in the $\mathcal{C}$-subgame when types $x \in [x_{-t-1}, x_{-t+1}]$ will next concede:

$$P^{\mathcal{C}}(x_{-t-1}, y, x_{-t+1}) = \int_{x_{-t-1}}^{x_{-t+1}} (1 - 2\pi(y,x) - \kappa) f(x|y)dx - \int_{-\infty}^{x_{-t-1}} 2\kappa f(x|y)dx \qquad (6)$$

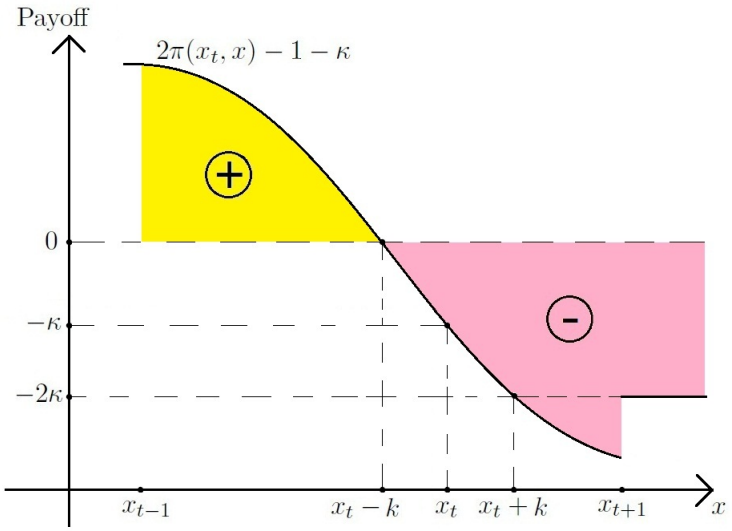


Figure 2: **The Propensity to Hold Out.** This is the area under the above net benefit curve, weighted by the conditional type density $f(x|y)$.

The indifference conditions of cutoff types $x_t$ and $x_{-t}$ are then given by:

$$P^{\mathcal{A}}(x_{t-1}, x_t, x_{t+1}) = 0 = P^{\mathcal{C}}(x_{-t-1}, x_{-t}, x_{-t+1}) \tag{7}$$

To analyze Moritz' initial cutoff type $x_0$ in a sincere agreeable responsive strategy profile, consider his initial propensity to vote $\mathcal{A}$ given Lones' next cutoffs $x_1$ and $x_{-1}$:

$$P^0(x_{-1}, m, x_1) = \int_{-\infty}^{x_{-1}} \kappa f(\ell|m)d\ell + \int_{x_{-1}}^{x_1} (1-2\pi(\ell, m))f(\ell|m)d\ell + \int_{x_1}^{\infty} -\kappa f(\ell|m)d\ell \tag{8}$$

This expression reflects three type ranges. For any type $\ell \leq x_{-1}$, Lones agrees immediately in the $\mathcal{A}$-subgame, but holds out in period one of the $\mathcal{C}$-subgame. Then, proposing $\mathcal{A}$ rather than $\mathcal{C}$ leads to the same verdict, acquittal, but reduces waiting costs by $\kappa$. If Lones agrees immediately in both subgames, so for types $\ell \in (x_{-1}, x_1)$, then Moritz' initial proposal becomes the verdict. In this case, proposing $\mathcal{A}$ rather than $\mathcal{C}$ decreases decision costs by $1 - 2\pi(\ell, m)$. Lastly, if Lones agrees immediately in the $\mathcal{C}$-subgame, but holds out in period one of the $\mathcal{A}$-subgame, so for types $\ell \geq x_1$, then proposing $\mathcal{A}$ rather than $\mathcal{C}$ leads to the same verdict, conviction, but raises waiting costs by $\kappa$. Moritz's type $m = x_0$ is indifferent between initially proposing $\mathcal{C}$ or $\mathcal{A}$ (and conceding immediately if Lones does not agree) if:

$$P^0(x_{-1}, x_0, x_1) = 0 \tag{9}$$

We next show that the indifference conditions are essentially necessary and sufficient for equilibrium. Call the cutoffs *tight* if whenever all types of one juror concede in some period, all remaining types of the other juror thereafter hold out forever. For example, if $x_t = \infty$ at some odd period $t$, then set $x_{t'} = x_{t-1}$ for all even $t' > t$.

**Theorem 1 (Equilibrium Characterization)** *Any sincere agreeable responsive equilibrium is equivalent to tight cutoffs* $(x_t)$ *that obey monotonicity* (4)*, indifference* (7) *and* (9) *when finite, and* $|x_0| < \infty$*. Conversely, any such cutoffs define a sincere agreeable responsive equilibrium.*

We now explore an important way in which the jurors' equilibrium cutoffs interact with one another. Reconsider the indifference condition (7) of cutoff type $x_t$. This type balances the possibility of improving the verdict against the associated waiting cost; thus, he must surely wish to hold out against the weakest remaining type of his colleague, $x_{t-1}$. As seen in Figure 2, this implies:

$$x_t - x_{t-1} > k \tag{10}$$

Thus, equilibrium cutoff types $(x_t)$ zick-zack around the 45-degree line in $(\ell, m)$-space.

For insight into the strategic nature of our problem, we now compare it to a natural embedded decision problem. The *private posterior* answers the *dictator's problem*: how one non-omniscient juror would vote if he were suddenly given dictatorial power and asked to choose the verdict unilaterally. This differs from our *debater's problem*, where the refusal to concede is not final, and arguments cost $\kappa$ per period. Since the debater must pay a deliberation cost and the dictator need not, one might think that the debater is more willing to concede than the dictator. We now argue that for small waiting costs, the opposite occurs: Namely, one may sometimes act as a *devil's advocate*, arguing for a verdict despite not believing it and not wishing that one's vote be the final word. For seconding a proposal ends the game, while holding out retains the option value of conceding later based on additional information about the other juror's type.

To see that devil's advocate types exist, say for Moritz, let $\bar{x}_t$ be his indifference cutoff in the dictator's problem in period $t$ when he has just learnt that Lones' type lies above $x_{t-1}$; that is, Moritz thinks that Lones is right and the defendant is guilty iff $m < \bar{x}_t$. Type $m$ is thus a devil's advocate in period $t$ if $x_t < m < \bar{x}_t$. The weakest remaining types of Moritz, who lie below $x_{t-1}$, know for sure that Lones is right;[8] so by

[8]In contrast to a standard war of attrition, these types concede in period $t$ not because their cost-benefit consideration turns negative, but because they are convinced that Lones is right.

continuity, the indifferent type satisfies $\bar{x}_t > x_{t-1}$. But devil's advocates exist when we have more strongly that $\bar{x}_t > x_t$. By §6, this holds in the efficient equilibrium for small enough waiting costs $\kappa > 0$; for $x_t - x_{t-1}$ vanishes as $\kappa \to 0$, whereas the dictator's threshold $\bar{x}_t$ depends only on $x_{t-1}$ but not the arguing costs $\kappa$.

For a different perspective on the argument, consider the belief of an uninformed observer. For instance, suppose the president is watching two informed members of his security council debate over whether to get bin Laden. Should he be worried about acting on the latest twist in the debate, if he is aware of the possibility of either party playing devil's advocate? He must predicate his behavior on the *public posterior* — that is, the probability of guilt conditional on history and equilibrium strategies. In informational herding models, this belief always favors the action just taken. The same result holds in our model with symmetric types. For suppose that Moritz has just held out and argued for acquittal in period $t$. Then the public posterior conditions on $\ell \geq x_{t-1}$ and $m \geq x_t$; by the zick-zack property we know $x_{t-1} < x_t$, so that with symmetric types, the public posterior favors acquittal.[9]

# 4 Deferential and Communicative Equilibria

We now show that there are infinitely many equilibria that differ in how much jurors invest into information aggregation. For motivation, assume that Moritz initially votes for acquittal if $m > 0$ and for conviction if $m < 0$, and plans to insist on this initial vote forever after. Then Lones cannot affect the verdict, and concedes immediately. So this strategy profile constitutes a Bayes-Nash equilibrium; with a judicious choice of off-path beliefs, this can be rationalized as a sequential equilibrium. Since Lones defers to Moritz immediately, we call it the *deferential* equilibrium. It intuitively corresponds to an asymmetric outcome of a standard war of attrition.

One might think of these instead as "insistent equilibria" — namely, where all types of Moritiz "dig in their heels" and insist on an outcome. But deferential is a broader notion. For even if not all types of Moritz insist, so few may concede that all types of Lones optimally defer in period one. Our terminology thus focuses on Lones' deferential behavior rather than Moritz' insistent behavior.

More generally, in a $(\sigma, \tau)$-*stopping profile* communication terminates by period $\sigma$ of the $\mathcal{A}$-subgame and period $\tau$ of the $\mathcal{C}$-subgame, where these *drop-dead dates* $\sigma$ and $\tau$ can be finite or infinite. They might be enforced by either protocol or regulation,

[9] In contrast, with asymmetric types, say if Lones' type distribution puts higher weight in the tails than Moritz', the public posterior may well favor conviction after Moritz' proposal to acquit.

such as a cloture rule in a parliament. A strategy profile is $(\sigma, \tau)$-*deferential* if it is an $(\sigma, \tau)$-stopping profile, where $(\sigma, \tau)$ is minimal and $\sigma < \infty$ or $\tau < \infty$ or both. Finally, a strategy profile is *communicative* if every period in either subgame is reached with positive probability. Thus, any strategy profile is either communicative or $(\sigma, \tau)$-deferential for some $(\sigma, \tau)$.

The characterization in Theorem 1 gives a possibly infinite set of equations (7) and (9) to verify equilibrium. Rather than pursue this involved route, we instead open a convenient backdoor.[10] As is well-known, Crawford and Haller (1990), common interest games satisfy a second welfare theorem: Any planner's solution for the game is also an equilibrium, since any deviation from an efficient strategy profile raises joint costs, and thus own costs. As efficient outcomes are easier to analyze than equilibria, this allows us to deduce equilibrium existence and a Pareto ranking among equilibria. So motivated, a sincere agreeable strategy profile is $(\sigma, \tau)$-*constrained efficient* if it maximizes *ex ante* expected payoffs in the class of sincere agreeable $(\sigma, \tau)$-stopping profiles.

**Theorem 2 (Existence)** *A $(\sigma, \tau)$-constrained efficient strategy profile exists for any $(\sigma, \tau)$. For $\sigma = \tau = \infty$, this strategy profile is a communicative equilibrium. For $\sigma$ or $\tau$ finite, this profile is equivalent to a $(\sigma, \tau)$-deferential equilibrium.*

We next study the dynamic stability of these equilibria.

**Theorem 3 (Sequentiality)** *The communicative equilibrium is a sequential equilibrium. Any $(\sigma, \tau)$-deferential equilibrium is equivalent to a sequential equilibrium.*

*Proof:* In a communicative equilibrium all information sets are reached on path, so that Bayes' rule determines beliefs at these information sets, and the resulting assessment is automatically a sequential equilibrium.

Now consider the $(\sigma, \tau)$-deferential equilibrium, say with $\sigma, \tau$ finite and odd; up to equivalence we can assume that no remaining type of Moritz concedes off-path, when Lones unexpectedly holds out in period $\tau$. Consistency in sequential equilibrium does not restrict Moritz' beliefs; for there exists a sequence of completely mixed strategies such that the sequence of beliefs induced by Bayes' rule converges to any desired beliefs over Lones' types. In particular, it is consistent for Moritz' type $m$ to believe that Lones' type satisfies $\ell < m-k$ with probability one, i.e., Moritz interprets Lones' failure to concede as a tremble of weak types. This rationalizes Moritz' insisting behavior. For

[10] We will come back to equations (7) and (9) when discussing equilibrium uniqueness, Theorem 5.

given such beliefs, and anticipating that Lones is about to concede, it is sequentially rational for Moritz to hold out. □

Deferential equilibria represent a communication failure. Strong types of Lones defer not because they are convinced that Moritz' type is stronger, but because they cannot bring Moritz to concede. This could not happen if Moritz was required to interpret off-path behavior of Lones as a signal of strength, rather than as a mistake. So, inspired by Cho (1987)'s definition for finite games, let us say that a sequential equilibrium in our game obeys *forward induction* if a player, say Moritz, who observes a deviation from the equilibrium path, must assign probability zero to types of Lones for whom the observed deviation is not sequentially rational — given any conjecture about future play and equilibrium beliefs over Moritz' types.

**Theorem 4 (Forward Induction)** *The communicative equilibrium satisfies forward induction, but no* $(\sigma, \tau)$*-deferential equilibrium satisfies forward induction.*

*Proof:* Forward induction has no bite in a communicative equilibrium, as every action node is reached on the equilibrium path. Next, consider the $(\sigma, \tau)$-deferential equilibrium, with $\sigma, \tau$ finite and odd, and assume that no remaining type of Moritz concedes off-path.[11] Forward induction forces Moritz to infer that Lones' type is strong if Lones unexpectedly deviates by holding out in period $\tau$. For Lones' equilibrium beliefs in period $\tau$ assign probability one that Moritz' type obeys $m > x_{\tau-1}$. But then holding out is not sequentially rational for Lones' types $\ell \leq x_{\tau-1}$ given any conjecture about future play; for such types are convinced that acquittal is the best verdict, *and* can immediately achieve this verdict with zero delay by conceding. Conversely, holding out is sequentially rational for sufficiently high types of Lones if they conjecture that Moritz changes his future behavior. So in period $\tau+1$, forward induction forces Moritz to believe $\ell > x_{\tau-1}$ almost surely. But then Moritz' weakest remaining type $m = x_{\tau-1}$ is convinced that conviction is the smart verdict, and insisting on acquittal is not sequentially rational. So this deferential equilibrium violates forward induction. □

Theorem 2 establishes that there are infinitely many classes of equilibria, indexed by $(\sigma, \tau)$. We now show that within each class, equilibrium is unique.

**Theorem 5 (Uniqueness)** *The communicative equilibrium is unique. For any* $(\sigma, \tau)$ *the* $(\sigma, \tau)$*-deferential equilibrium is unique up to equivalence.*

[11] In the appendix we show that any $(\sigma, \tau)$-deferential equilibrium with different off-path strategies, $x_{t'} > x_{\tau-1}$ for some $t' > \tau - 1$, does not satisfy forward induction either.

For insights into our argument, let's revisit our incentive equations (7) and (9), which constitute a second-order difference equation on the cutoff sequence $(x_t)$. The equilibria only differ by their terminal conditions: The $(\sigma, \tau)$-deferential equilibrium must satisfy $x_{-\sigma} = -\infty$ and $x_\tau = \infty$; the communicative equilibrium must satisfy $x_{-t} \to -\infty$ and $x_t \to \infty$. Thus, equation counting already suggests uniqueness.

For a tighter intuition, note that the propensity to hold out $P^{\mathcal{A}}(x_{t-1}, x_t, x_{t+1})$ falls in its first and third arguments: For the positive area in Figure 2 decreases in $x_{t-1}$, while the negative area increases in $x_{t+1}$. Thus, we can write (7) forwardly as $x_{t+1} = \phi(x_{t-1}, x_t)$, where $\phi_1 < 0$. Moreover, an identical increment in $x_{t-1}, x_t$ and $x_{t+1}$ in Figure 2 shifts the density left towards the positive area in the MLRP sense by Lemma 1, but leaves the integrand in (5) unchanged. Thus, the propensity tips positive, and $x_{t+1}$ must further increase to restore balance. In other words, $\phi_1 + \phi_2 > 1$.

Next, for a fixed anchor $x_0$, any choice of seed $x_1$ determines the entire cutoff sequence $x_2, x_3, \ldots$ by repeated application of the *shooting function* $\phi$. Indeed, we can show inductively that $x_t$ increases in $x_1$, for all $t > 1$. For if $dx_t \geq dx_{t-1}$, then:

$$dx_{t+1} = \phi_1 dx_{t-1} + \phi_2 dx_t = \phi_1(dx_{t-1} - dx_t) + (\phi_2 + \phi_1)dx_t > dx_t. \tag{11}$$

Thus, for *finite* $\tau$, there is a unique seed $x_1$ with $x_\tau = \infty$ and uniqueness of the $(\sigma, \tau)$-deferential equilibrium follows.

Uniqueness of the communicative equilibrium is harder to prove as two different cutoff sequences might explode at different rates. The proof strengthens the monotonicity argument as follows. The type density $f(y|x)$ eventually falls in $y$, so that the propensity $P^{\mathcal{A}}(x_{t-1}, x_t, x_{t+1})$ is more responsive to its first than to its third argument. Thus, the derivative $\phi_1$ is not only negative, but more strongly satisfies $\phi_1 < -(1 + \epsilon)$ for some $\epsilon > 0$. And we can see from (11) that $dx_{t+1} - dx_t$ then grows geometrically:

$$dx_{t+1} - dx_t = \phi_1(dx_{t-1} - dx_t) + (\phi_2 + \phi_1 - 1)dx_t > (1 + \epsilon)(dx_t - dx_{t-1}) \tag{12}$$

This implies that the dynamical system defined by the shooting function strictly "fans out" (as seen in Figure 3). A unique equilibrium intuitively arises just as it does for saddle point stable equilibria in growth theory. If $(x_t)$ is a communicative equilibrium, then any higher seed $x'_1 > x_1$ leads to diverging step-sizes $x'_{t+1} - x'_t = (x_{t+1} - x_t) + (dx_{t+1} - dx_t)$. This is inconsistent with a communicative equilibrium, as the domain of $\phi(\cdot, x_t)$ is bounded below; for if the positive area is too large in Figure 2, the propensity to hold out is positive even as $x_{t+1} \to \infty$.

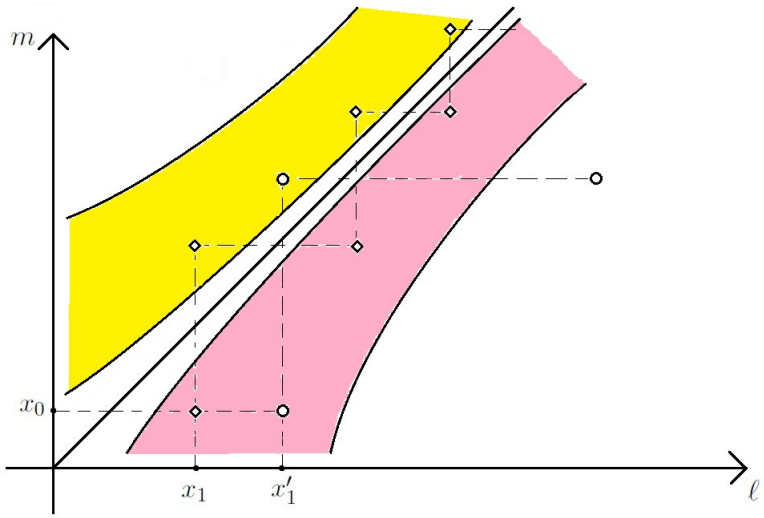


Figure 3: **Equilibrium Cutoffs as a Dynamical System.** Together with anchor $x_0$, equilibrium seed $x_1$ and shooting function $\phi(x_{t-1}, x_t) = x_{t+1}$ define sequence of diamonds $(x_1, x_2), (x_3, x_2), \cdots$. The sequence of circles, defined by the alternative seed $x_1'$, fans out and eventually leaves the domain of $\phi(\cdot, x_t)$ (shaded).

# 5 A Puzzle about Conversational Efficiency

We now return to the perspective of the uninformed observer. As the debate transpires without reaching agreement, the jurors must hold strong but countervailing information. What does he infer about their joint information? Do the strong opinions imply that a greater signal imbalance is more likely? Or might disagreement indicate the opposite, that we should expect a more moot conversation?

The absolute value of the type differential $|\delta| = |\ell - m|$ is a good ordinal measure for the payoff relevant information of the types since the difference in decision costs of the two verdicts equals $|2\pi(\ell, m) - 1| = |(e^\delta - 1)/(e^\delta + 1)| = (e^{|\delta|} - 1)/(e^{|\delta|} + 1)$. We thus adopt $|\delta|$ as a measure of mootness of the learnable information $(\ell, m)$. Say that conversations *grow moot after Moritz* in an equilibrium $(x_t)$ if, given that Lones argues for the correct verdict and Moritz holds out past period $t$, i.e. $\delta > 0$ and $m > x_t$, the distribution of $\delta$ decreases along the sequence $t, t+2, t+4, \ldots$ in the MLRP order; for then $|\delta|$ also falls in the weaker FOSD order, and thus so do the expected losses from taking the wrong verdict. We analogously define when conversations *grow moot after Lones*. If both mootness assertions hold, then we say that conversations *grow moot*.

**Proposition 1 (Das Moot)** *In any equilibrium, conversations grow moot.*

*Proof:* The density of $\delta$ conditional on $\delta > 0$ and $m > \underline{m}$ is given by

$$\int_{\underline{m}}^{\infty} h(m+\delta, m)dm = \int_{\underline{m}}^{\infty} f(m+\delta|m)f(m)dm = \int f(m+\delta|m)f(m)\mathbb{I}_{\{m>\underline{m}\}}dm \quad (13)$$

By Lemma 1, $f(m+\delta|m)$ is log-supermodular in $m$ and $-\delta$. The indicator function $\mathbb{I}_{\{m>\underline{m}\}}$ is log-supermodular in $m$ and $\underline{m}$. So by Proposition 3.2 in Karlin-Rinott, (13) is log-supermodular in $\underline{m}$ and $-\delta$. Then the distribution of $\delta$ is higher in the MLRP order at $x_t$ than at $x_{t+2} > x_t$. □

The proof noticeably makes no appeal to the equilibrium magnitudes of these thresholds but simply owes to the following statistical property of log-concave densities: If jurors have opposing types, then when Moritz' type is stronger than Lones', the differential is more likely smaller the stronger is Lones' type. So when comparing conversations across type realizations, our model predicts a negative correlation between the realized waiting costs and the value of the aggregated information, $(e^{|\delta|}-1)/(e^{|\delta|}+1)$. In other words, our conversational war of attrition is subject to *diminishing total returns*. In particular, the event when players have strong and opposing types leads to unboundedly bad payoff outcomes in the communicative equilibrium. Importantly, long conversations are not exceedingly rare events, as we now argue.

**Proposition 2 (Lengthy Conversations)** *In the communicative equilibrium, the hazard rate of conversation ending is bounded if and only if $f/(1-F)$ is bounded.*

*Proof:* The conversation ends with a hazard rate $(F(x_{t+2}) - F(x_t))/(1 - F(x_t)) = \int_{x_t}^{x_{t+2}} f(x)/(1 - F(x_t))dx$. In the communicative equilibrium, $x_{t+2} - x_t \geq 2k$, via the zick-zack property (10). But it is also bounded above by $x_{t+2} - x_t \leq x_2 - x_0 < \infty$, since $x_{t+1} - x_t$ strictly falls, by Lemma 7 in Appendix B.6. To wit, the conversation ends at an exploding hazard rate iff the type distribution hazard rate $f/(1-F)$ explodes. □

For the earlier distribution $f(x) = e^x/(1+e^x)^2$, the hazard rate is bounded and thus so is the hazard rate of conversation ending in the communicative equilibrium. On the other hand, if the log-likelihood ratio had a Guassian distribution, then the hazard rate would explode, and conversations would therefore quickly conclude.

If conversations grow moot, does this argue in favor of a cloture rule that truncates them when they have persisted long enough? We now argue not. Impose the standard partial order on equilibria based on their drop-dead dates $(\sigma, \tau)$.[12,13] Then:

[12] By convention, the communicative equilibrium has drop-dead dates $(\infty, \infty)$.

[13] For symmetric deferential equilibria, where $\sigma = \tau$, the weak requirement of a later drop-dead date

**Proposition 3 (Pareto-ranking)** *Equilibria are strictly Pareto-ranked by length, and in particular, the communicative equilibrium is the most efficient one.*

*Proof:* A constrained efficient strategy profile is an equilibrium, by Theorem 2, and there are no other equilibria, by Theorem 5. Relaxing the constraint that the conversation end by period $\tau$ obviously weakly raises welfare. But since constraints bind in a deferential equilibrium by Theorem 2, a strict Pareto-ranking of equilibria emerges. □

While paradoxical, the ex-ante optimal strategy of many related sequential decision problems feature ex-post total diminishing returns as well. For example, consider a toy version of Wald (1947)'s model of sequential Bayesian analysis that also shares the layered uncertainty of our strategic setting, in which a signal $X$ about the true state $\theta$ is learnable but $\theta$ is not. Specifically, assume that Nature draws both a state $\theta \in \{\mathcal{I}, \mathcal{G}\}$ with equal chances, and a signal $X \in [0,1]$ with $X = \Pr(\theta = \mathcal{G})$. Judge Lones knows neither $\theta$ nor $X$, but entertains a prior that $X$ is distributed uniformly on $[0,1]$, and thus that $Pr(\theta = \mathcal{G}) = 1/2$. Then $|X - 1/2|$ measures the mootness of the learnable signal $X$, with the realization $X = 1/2$ leaving him indifferent between the two verdicts. Assume that judge Lones can learn about $X$ through sequential Bernoulli trials at cost $\kappa > 0$ and can take the verdict at any time. Then there is a stopping set, invariant to $X$, and when his signal draws land in that set, he quits experimenting and decides. As $X$ grows more moot — by approaching $1/2$— one can show that Lones stochastically lands in the stopping set at a later time. All told, total investment negatively correlates with the value of the information for the decision, just as in our strategic setting.

More simply, consider a toy version of Weitzman (1979) optimal stopping problem with uncertain alternative prizes: Indiana Lones faces boxes $i = 1, \ldots, n$; independently across boxes, box $i$ contains a treasure of size $x_i$ with probability $p_i$, where $x_1 > x_2 > \cdots$. Lones opens the boxes sequentially at cost $\kappa$ and can only keep one treasure. Optimally, he opens the boxes in order of decreasing Gittins indices $w_i = x_i - \kappa/p_i$. If the chance $p_i$ is constant in $i$, then he opens the boxes in their natural order $1, 2, \ldots$. In this case, the total realized returns $x_i$ decrease over time.[14] Long searches correspond to the failure realizations of the early boxes. Similarly, jurors Lones and Moritz first search their joint types $(\ell, m)$ in parts of the type space where the aggregated information is most valuable, and repeated failure indicates that $(\ell, m)$ is moot.

$\tau' > \tau$ implies an *accordion property*: All cutoffs decrease, $x'_t < x_t$, so that the longer equilibrium conversation is more refined and takes more periods to agreement for *every type profile* $(\ell, m)$; this property follows from the proof of Lemma 8 in Appendix B.6.

[14] Weitzman gives examples in which the value of the option to learn about unopened boxes can lead one to open boxes in an order opposite to the static best. For instance, the index $w_i$ may decrease even if the expected values $p_i x_i$ slightly increases in $i$.

# 6 The Cheap Talk Limit

We now turn to a natural comparative static of the communicative equilibrium: What happens as the period length $\kappa$ falls, and thus talking grows cheaper. We focus in particular on the "cheap talk" limit, since that is a focal point in the theory literature.

**Proposition 4 (Cheaper Talk)** *For any types $(\ell, m)$, the time to agreement in the communicative equilibrium falls in $\kappa$.*[15]

In a standard war of attrition, the time to agreement in the focal, longest equilibrium is bounded away from zero as the period length vanishes; that is, the increased number of periods to agreement is asymptotically inverse to the period length. We argue instead that — just as happens in the Coase Conjecture limit — the expected time to agreement in the communicative equilibrium vanishes in the limit $\kappa \downarrow 0$. More strongly, we shall characterize how fast waiting costs and decision costs vanish.

For some context, let us consider a decision theory benchmark close in spirit to our model. Imagine judge Moritz waiting on a "smoking gun" signal. In the innocent state, no such signal comes, while in the guilty state, its arrival rate is $\alpha > 0$ per period, and the period length is $\kappa > 0$. In this world, decision costs optimally vanish like $\kappa$ near zero, while waiting time grows like $\log(1/\kappa)$.[16] Thus, costs vanish like $\kappa \log(1/\kappa)$, and so the ratio of decision costs to waiting costs vanishes.

Just like judge Moritz, juror Moritz waits on the arrival of a decisive event — a smoking gun here, versus Lones' concession to convict in our model. However, the informativeness of the absence of a smoking gun is exogenous while the probability of Moritz conceding in any given period vanishes as $\kappa \to 0$, and so the absence of this event becomes uninformative. Thus juror Moritz learns slower than judge Moritz, and costs accordingly vanish at a slower rate than $\kappa \log(1/\kappa)$. In fact, we argue below that the jurors' costs vanish at rate $\kappa^{2/3}$ for small $\kappa$.

Formally, let $v^*(\ell, m)$ and $v_\kappa(\ell, m)$ denote the ex-post correct verdict and the realized verdict in the efficient equilibrium with cost $\kappa > 0$. The relative cost of ex-post

[15] This is also true of the unique symmetric $(\tau, \tau)$-deferential equilibrium, for any $\tau < \infty$.

[16] Absent a smoking gun, the log-likelihood ratio of the judge's Bayesian posterior on innocence evolves according to $m_{t+1} = m_t - \log(1-\alpha)$. When the judge quits his search for the smoking gun and acquits the defendant his decision costs equal the remaining probability of guilt, $p(m) := 1/(1+e^m) \approx e^{-m}$ as $m \to \infty$. The marginal benefit of searching one more period is $p(m) - p(m - \log(1-\alpha)) \approx \alpha e^{-m}$. Thus, the judge optimally quits when $\alpha e^{-m} < \kappa$ or, equivalently, $m > \log(\alpha/\kappa)$. Therefore, decision costs vanish at rate $\kappa$ while waiting costs vanish at rate $\kappa \log(1/\kappa)$, total costs vanish at rate $\kappa \log(1/\kappa)$ and the ratio of decision costs to waiting costs converges to zero as $\kappa \to 0$.

correct and incorrect verdict equals $|\pi(\ell, m) - (1 - \pi(\ell, m))|$,[17] so we define expected *decision costs* as

$$d(\kappa) = \int_{v_\kappa(\ell,m) \neq v^*(\ell,m)} |2\pi(\ell, m) - 1| h(\ell, m) d\ell dm$$

Let $T(\kappa)$ be the expected number of periods to agreement. So expected waiting costs are $\kappa T(\kappa)$ and total costs $c(\kappa) = d(\kappa) + \kappa T(\kappa)$.

**Proposition 5 (Very Cheap Talk)** *The total costs $c(\kappa)$ are of order $\Theta(\kappa^{2/3})$ for vanishing $\kappa$.*[18]

The proof of Proposition 5 analyzes the cutoff strategy profiles with constant *step-length*, $x_{t+1} - x_t \equiv \Delta$ for all $t$, and shows that the desired bounds are asymptotically achieved by setting $\Delta = \kappa^{1/3}$. To understand why this step-length is asymptotically optimal, consider the dynamic programming problem of a planner who has observed cutoffs up to $x_{t-1}$, and must instruct Lones which cutoff $x_t$ to employ. He faces three costs: (1) incremental waiting costs, (2) decision costs in the event that Lones wrongfully concedes to acquit, and (3) continuation costs, when he holds out.

Decision costs next period — which arise for $(\ell, m)$ with $x_{t-1} < m < \ell < x_t$ — equal an integral over a triangle of side $\Delta$; thus, they are approximately $\alpha\Delta^3$, for some constant $\alpha > 0$. Likewise, the chance that Lones concedes in the next step approximately equals $\beta\Delta$, for some constant $\beta > 0$. Thus, the expected continuation costs are approximately $(1 - \beta\Delta)c(\kappa)$. This yields the approximate planner's Bellman equation:

$$c(\kappa) \approx \kappa + \alpha\Delta^3 + (1 - \beta\Delta)c(\kappa) \qquad \Rightarrow \qquad c(\kappa) \approx (\kappa/\Delta + \alpha\Delta^2)/\beta$$

To minimize $c(\kappa)$, the planner chooses $\Delta = \kappa^{1/3}/(2\alpha)^{1/3}$.

This Bellman logic leads to a surprising corollary. For small waiting costs $\kappa$, the ratio of decision costs to waiting costs is $\alpha\Delta^3/\kappa \approx 1/2$ in ever single period.[19] Unlike for judge Moritz, this ratio no longer vanishes, because juror Moritz faces juror Lones, whose concession rate is vanishing.

[17] This term is net of the statistical decision costs $\min\{\pi(\ell, m), 1 - \pi(\ell, m)\}$ with respect to the unlearnable state $\theta$. It corresponds to the *full information gap* of Moscarini and Smith (2002) with respect to $(\ell, m)$.

[18] This means that $\underline{\theta}\kappa^{2/3} \leq c(\kappa) \leq \bar{\theta}\kappa^{2/3}$ for constants $0 < \underline{\theta} \leq \bar{\theta} < \infty$ and all small enough $\kappa > 0$.

[19] This can be made rigorous, but the asymptotic proof is somewhat lengthy.

# 7 The Conclusion: Two Roads Not Taken

We have explored the implications of coarse and entirely dispositive communication. As with any theory, these model pillars are overly simple assumptions, but still both merit defense. We now address these in sequence.

## 7.1 Why is communication coarse?

If jurors share identical preferences, one might ask why they don't just lay their signals on the table and optimally combine them? Yet anyone who has been in debates with like-minded peers is keenly aware that this does not happen. One simple reason is that — just as is often argued for failures of transitive preferences, that rational individuals have to introspect to deduce their true preferences — a Bayesian may not be fully aware of his information. If so, then he might well engage in an introspective process of signal discovery. A simple theory of *mental experiments* then captures the difficulty of this process consistent with our model.

Suppose that each juror can sequentially ask himself whether his belief is stronger or weaker than a threshold of his choice. A strategy is then a sequence of thresholds. Compared to the jurors in our model, who know their types, these jurors have strictly less information and can achieve weakly lower payoffs. But as best responses in our model are in cutoff strategies, they can be emulated by jurors who have to learn their types through mental experiments. Best response strategies and *equilibria thus coincide for the two models*. In other words, we can understand the coarse communication as simply the outcome of an introspective learning exercise transpiring period by period.

## 7.2 Just Talk

To reach a verdict, communication must eventually become dispositive. Our model takes this idea to the extreme by assuming that the jurors' proposals are dispositive from the very beginning of the game; they cannot say "my signal indicates guilt but let's not jump to conclusions yet". Consider now a model with non-dispositive communication. For instance, assume that in any period a juror can not only move to acquit or convict the defendant, but also humbly suggest guilt or innocence. The game ends after two consecutive motions for the same verdict.

Best response strategies in this model are difficult to characterize because each juror's Bellman equation depends on the interval of his colleague's remaining types.[20]

[20]In contrast, our analysis does not require dynamic programming and relies instead on the indif-

Intuitively, humble suggestions leave a lot of flexibility for encoding one's type. For example, Moritz could adopt a contrarian strategy where low types $m$ (that indicate guilt) move to acquit, higher types suggest innocence, yet higher types suggest guilt, and the highest types move to convict. If Lones interprets this strategy correctly, it may well be part of a best response. In contrast to our model, where such contrarian behavior can arise only in period zero, it could go on for a long time here.

Nonetheless, main insights of our paper carry over to this less tractable model. Importantly, all equilibria in our model remain equilibria in this model if jurors ignore humble suggestions by their colleague and just repeat their last motion; for then humble suggestions are just a waste of time. Thus, there are still deferential equilibria, where the conversation terminates in some period $\tau$ with probability one, and communicative equilibria with unbounded length. And for the same reason as in our model, deferential equilibria are Pareto-dominated and violate forward induction. However, there are now also other equilibria where jurors make use of humble suggestions. Back-of-the envelope calculations similar to the ones after Proposition 5 and in footnote 16 show that total costs decrease at rate $\kappa \log 1/\kappa$; thus, non-dispositive communication increases welfare. While a full analysis of this richer model is outside the scope of this paper, this discussion indicates that many of our arguments and insights can be extended to models that allow for non-dispositive communication.

# A Appendix: Contrarian Contradictory Equilibria

Our analysis has focused on sincere agreeable equilibria. Assume to the contrary that Moritz is a contrarian who initially proposes $\mathcal{A}$ if his signal indicates guilt, say $m < x_0 = 0$, and $\mathcal{C}$ if his signal indicates innocence. To see that contrarianism can arise in equilibrium, suppose further that Moritz defers to Lones in period two. Thus, Lones calls the verdict in period one and must be careful to second Moritz only if his type strongly favors this verdict; for after all an $\mathcal{A}$ vote by Moritz indicates guilt. Formally, Lones' first cutoff in the $\mathcal{A}$-subgame must satisfy $x_1 < 0$, and similarly $x_{-1} > 0$. Thus, Lones' best response to Moritz' contrarian strategy is contradictory, with types $\ell \in [x_1, x_{-1}]$ contradicting either one of Moritz' initial proposals. Anticipating that his initial proposal will likely be overturned, and assuming additionally that Lones will insist on his verdict after period one, Moritz' contrarian strategy is indeed optimal. The outcome of this equilibrium is illustrated in Figure 4.

ference conditions (7) and (9); these conditions suffice because our game-form is a war of attrition where players initially commit to a concession period.

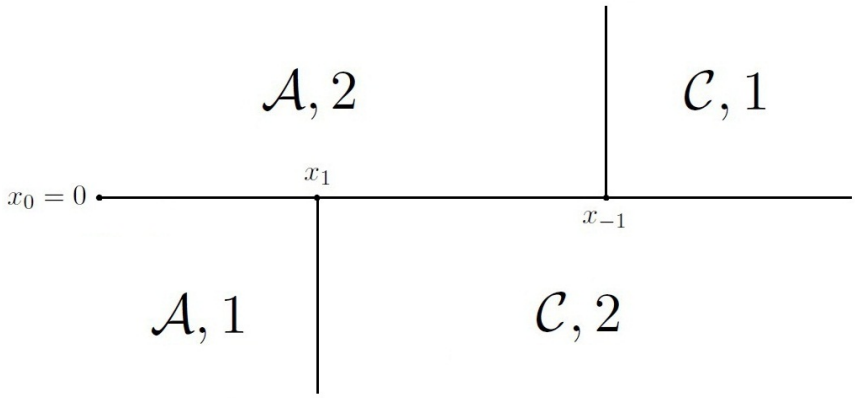


Figure 4: **Outcomes of a Contrarian Contradictory Equilibrium.**

Such discordance between the literal semantics of a proposal like "Acquit" and its equilibrium interpretation is common in the cheap-talk literature, e.g., Crawford and Sobel (1982). In our game it can at most arise in period zero, for afterwards arguments are dispositive. But even in period zero, Moritz' equilibrium choice of arguments is not arbitrary: a sincere strategy allows Lones to agree and reach a verdict in period one in the likely case that the signals indicate the same verdict; the contrarian strategy forces Lones to contradict Moritz in period one and only reach the verdict in period two. More formally, the contrarian contradictory equilibrium is Pareto-dominated by the sincere agreeable $(2, 2)$-deferential equilibrium: Assume that Moritz argues sincerely in period zero and concedes for sure in period two, and that Lones plays according to cutoffs $x'_1 = x_{-1}$ and $x'_{-1} = x_1$. This strategy profile yields the same verdicts as the above contrarian contradictory equilibrium, but expected waiting costs are lower because Lones' types in the interval $(x_{-1}, x_1)$ now reach a verdict in period one rather than period two. If Lones best responds to Moritz' sincere strategy, rather than mimicking the cutoffs $x_{-1}, x_1$, costs decrease further.

There may exist other, $(\sigma, \tau)$-deferential contrarian contradictory equilibria where the conversation continues past period two. Moritz' behavior is contrarian only in period zero, but monotone thereafter by Lemma 2. Intuitively, such equilibria are subject to the same inefficiency, suggesting that the cost-minimizing strategy profile is indeed sincere and agreeable. We leave the proof of this conjecture for future work.

As contrarian contradictory equilibria differ from sincere agreeable equilibria qualitatively only in period zero, the main insights of our paper carry over to these equilibria. For instance, the above contrarian contradictory equilibrium violates forward induction and is Pareto-dominated by contrarian contradictory equilibria that last more periods. Also, conversations grow moot in contrarian contradictory equilibria.

# B Appendix: Omitted Proofs

## B.1 Properties of the Density Functions and Proof of Lemma 1

By the definition of our types $\ell$ and $m$, we have $\Gamma(\mathcal{G}|\ell) = \lambda = e^\ell/(1+e^\ell), \Gamma(\mathcal{I}|\ell) = 1-\lambda = 1/(1+e^\ell), \Gamma(\mathcal{I}|m) = \mu = e^m/(1+e^m)$, and $\Gamma(\mathcal{G}|m) = 1-\mu = 1/(1+e^m)$. The law of total probability implies $f(\ell) = \sum_\theta f^\theta(\ell)\Gamma(\theta)$, and the Radon-Nikodym derivative $f^\mathcal{G}(\ell)/f^\mathcal{I}(\ell) = \Gamma(\mathcal{G}|\ell)/\Gamma(\mathcal{I}|\ell)$ equals $\lambda/(1-\lambda) = e^\ell$. Altogether, we obtain:

$$\begin{aligned} f(\ell) &= \frac{1}{2}f^\mathcal{I}(\ell) + \frac{1}{2}f^\mathcal{G}(\ell) = \frac{1}{2}\left(1 + \frac{f^\mathcal{G}(\ell)}{f^\mathcal{I}(\ell)}\right) f^\mathcal{I}(\ell) = \frac{1+e^\ell}{2} f^\mathcal{I}(\ell) \\ f(\ell|m) &= f^\mathcal{I}(\ell)\Gamma(\mathcal{I}|m) + f^\mathcal{G}(\ell)\Gamma(\mathcal{G}|m) = \left(\frac{e^m}{1+e^m} + \frac{f^\mathcal{G}(\ell)/f^\mathcal{I}(\ell)}{1+e^m}\right) f^\mathcal{I}(\ell) = \frac{e^m+e^\ell}{1+e^m} f^\mathcal{I}(\ell) \end{aligned}$$

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

$$r(\ell, m) = \frac{2(e^\ell + e^m)}{(1+e^\ell)(1+e^m)}$$

As a product of terms just in $\ell$ or $m$, and $e^\ell + e^m$, this is log-submodular as $a+b$ is log-submodular: $(a'+b')(a+b) - (a'+b)(a+b') = -(a'-a)(b'-b) < 0$ if $a' > a, b' > b$.

We now turn to the proof of Lemma 1. As $r(\ell, m)$ is symmetric in $\ell$ and $m$ and they share the distribution $f$, we now adopt the generic variables $x, y$. We have $f(y+\delta|y) = f(y+\delta)r(y+\delta, y)$ and $f(y+\delta)$ is log-submodular by the log-concavity assumption on $f$. Also:

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

This is a product of log-submodular terms. Indeed, $1/(1+e^y e^\delta)$ 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 $a' > a, b' > b$.

## B.2 Sincere Agreeable Strategies: Proof of Lemma 3

The proof of Lemma 3 is in two steps. First we show that Moritz' best reply to a monotone and agreeable strategy of Lones is sincere, and then we argue that Lones' best reply to a monotone and sincere strategy of Moritz is agreeable.

Any agreeable monotone strategy of Lones entails thresholds ordered as follows:

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

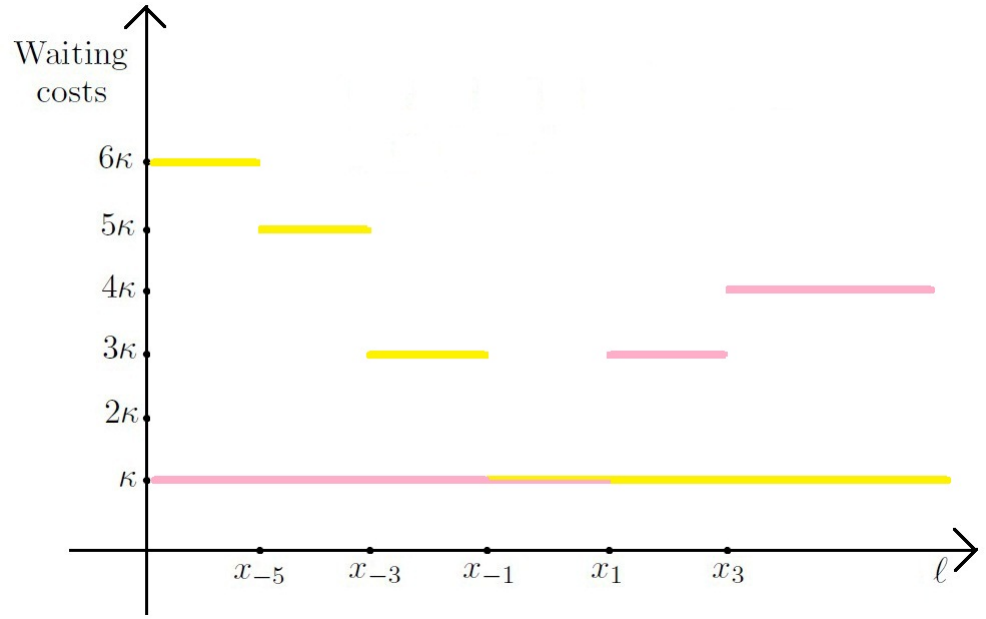


Figure 5: **Waiting Costs.** Moritz' waiting costs for $(\mathcal{A}, 4)$ (purple) and $(\mathcal{C}, 6)$ (yellow) as a function of Lones' type $\ell$. Indeed, assume Moritz plays $(\mathcal{A}, 4)$. If $\ell < x_1$, Lones immediately concedes and waiting costs are $\kappa$. If $\ell \in (x_1, x_3)$, Lones concedes in period 3 and waiting costs are $3\kappa$. If $\ell > x_3$, then Moritz concedes in period 4 and waiting costs are $4\kappa$.

Consider any two even periods $t, t' \geq 2$ and consider Moritz' strategy to initially propose $\mathcal{A}$ and concede in period $t$, $(\mathcal{A}, t)$, and his strategy to initially propose $\mathcal{C}$ and concede in period $t'$, $(\mathcal{C}, t')$.

**Claim 1** *The cost increment of* $(\mathcal{A}, t)$ *over* $(\mathcal{C}, t')$ *decreases in* $m$.

*Proof:* Figure 5 plots the waiting costs of $(\mathcal{A}, t)$ and $(\mathcal{C}, t')$ as a function of Lones' type $\ell$. Obviously, the difference $\Delta(\ell)$ in waiting costs ($(\mathcal{A}, t)$ less $(\mathcal{C}, t')$) weakly increases in $\ell$. Next, consider the change in decision costs. If $\ell \in (x_{-t'-1}, x_{t+1})$, the verdict changes, and the decision cost difference equals 1 in state $\mathcal{G}$ and $-1$ in state $\mathcal{I}$. Given state conditional densities $f^{\mathcal{I}}, f^{\mathcal{G}}$, Moritz' type $m$ expects that the cost differential equals:

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

Differentiating in $m$, and using $\Gamma_m(\mathcal{I}|m) = -\Gamma_m(\mathcal{G}|m)$:

$$G_m(\mathcal{G}|m)\left[\int_{x_{-t'-1}}^{x_{t+1}} (f^{\mathcal{G}}(\ell) + f^{\mathcal{I}}(\ell))d\ell + \int_{-\infty}^{\infty}\Delta(\ell)(f^{\mathcal{G}}(\ell) - f^{\mathcal{I}}(\ell))d\ell\right]$$

This is negative because $\Gamma(\mathcal{G}|m) = \frac{e^{-m}}{1+e^{-m}}$ decreases in $m$, the first integral is obviously positive, and the second integral is positive because $f^{\mathcal{G}}(\ell)$ MLRP-dominates $f^{\mathcal{I}}(\ell)$ (recall that $f^{\mathcal{G}}(\ell)/f^{\mathcal{I}}(\ell) = e^{\ell}$). This establishes Claim 1, and that Moritz' best response to an agreeable, monotone strategy is sincere.

Conversely, consider Lones' best response to a sincere strategy of Moritz. If Moritz is non-responsive, say he always initially argues $\mathcal{C}$, i.e. $x_0 = \infty$, then — up to equivalence — we can set Lones' first off-path cutoff $x_1$ equal to $\infty$, so that his strategy is agreeable. If Moritz is responsive with finite initial cutoff $x_0$ and Moritz initially argues $\mathcal{A}$, then Lones' types $\ell \leq x_0$ are convinced that $\mathcal{A}$ is the ex-post correct verdict and strictly prefer to concede. By continuity, Lones' best response must satisfy $x_1 > x_0$. Analogously, we get $x_{-1} < x_0$, as desired.

## B.3 Equilibrium Characterization: Proof of Theorem 1

First, consider sufficiency of the conditions. By definition, monotonicity (4) 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 (7) and (9), when finite.

We show that conceding in (odd) period $t+2$ of the $\mathcal{A}$-subgame is optimal for any type $\ell \in [x_t, x_{t+2}]$ of Lones. First, he weakly prefers this to stopping in period $t$ because type $x_t \leq \ell$ is indifferent between these strategies, and the payoff difference single-crosses in $\ell$, by Lemma 2. As $x_{t-2} \leq x_t \leq \ell$, the same argument shows that $\ell$ weakly prefers conceding in period $t$ to conceding in period $t-2$. By induction and transitivity, $\ell$ does not want to concede before period $t+2$. This single-crossing logic can also be used to 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 Moritz concedes after period $t'$, and so all types of Lones prefer conceding in period $t'$ to holding out longer — delay incurs waiting costs and does not change the verdict, as Moritz is set.

The analysis for Moritz is similar, except for the initial period. Indeed, proposing $\mathcal{A}$ initially and conceding in period $t+2$ is a best response for Moritz' types $m \in [x_t, x_{t+2}]$. In fact, to finish this argument, we also must exploit the assumption that type $x_0$ is indifferent between initially voting $\mathcal{A}$ and $\mathcal{C}$ (and conceding immediately if Lones disagrees) and that the cost difference between these two plans is decreasing, by Claim 1 (proof of Lemma 3).

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 that this fails, say, for odd $t$. 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.

Thus, any sincere agreeable equilibrium is characterized by cutoffs $(x_t)$ as in (4), where in each subgame there is a period $t$, possibly $\infty$, such that before $t$ the inequalities are strict, and in period $t+1$ the game terminates with probability 1. To fix ideas we assume WLOG that $t$ is odd and finite, so that $x_{t+1} = \infty$. Now it is easy to see that the finite equilibrium cutoff types must indeed be indifferent: In any odd period $t' < t$, Lones' equilibrium strategy has types just below $x_{t'}$ conceding in period $t'$, and types just above $x_{t'}$ conceding in period $t'+2$. By continuity of Lones' preferences in $\ell$, the cutoff type $x_{t'}$ must be indifferent. The same argument shows that in any even 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 response to concede in period $t+2$ for types just above $x_t$. Then, by continuity the equilibrium cutoff type $x_t$ is indifferent between conceding and holding out in period $t$ as required. Thus, (7) and (9) are necessary.

## B.4 Equilibrium Existence: Proof of Theorem 2

Step 1: A Constrained Efficient Outcome Exists. In looking for the constrained-efficient strategy profile we can restrict ourselves to monotone strategies represented by vectors of cutoff values $(x_t)$ because any best-response is in monotone strategies by Lemma 2. For this proof it is useful to transform the cutoffs from (unbounded) log-likelihood ratios to (compact) probabilities $\chi_t \equiv e^{x_t}/(1+e^{x_t}) \in [0,1]$. We can restrict attention to cutoff vectors $(\chi_t)$ where cutoffs in late periods have negligible payoff consequences, i.e., either $\lim \chi_{2t} = 1$ or $\lim \chi_{2t+1} = 1$; cutoff vectors that do not satisfy this property have expected total costs of $\infty$.

Any sequence of cutoff vectors approaching the cost infimum has a pointwise convergent subsequence $(\chi_t)^i$ by Tychonoff's theorem. The limit cutoff vector $(\chi_t)^*$ achieves

the cost minimum because cutoffs in late periods have negligible payoff consequences.

Step 2: A Constrained Efficient Outcome is a Bayesian Equilibrium. For the purpose of this proof, write Lones' and Moritz' strategies as $L$ and $M$. Consider an $(\sigma, \tau)$-constrained efficient strategy profile $(L, M)$. Assume WLOG that $\sigma$ and $\tau$ are odd, that is, Lones is deferential and concedes in periods $\sigma$ and $\tau$ of the subgames. Since the game almost surely ends in periods $\sigma$ and $\tau$, we can assume up to equivalence that no type of Moritz concedes after periods $\sigma - 1$ and $\tau - 1$.

To show that $(L, M)$ is indeed an equilibrium, consider any best reply $L'$ by Lones to $M$. Clearly, Lones almost surely concedes in periods $\sigma, \tau$, since Moritz will not change his mind afterwards. So $(L', M)$ is a $(\sigma, \tau)$-stopping profile.[21] For a contradiction, assume that a positive measure of Lones' types has strictly higher payoffs under $L'$ than under $L$. Then the common ex-ante payoffs of $(L', M)$ exceed those of $(L, M)$, contradicting constrained efficiency of $(L, M)$. By the same argument, any best response of Moritz to $L$ must coincide with $M$ up to periods $\sigma$ and $\tau$. Moritz' behavior after these periods is payoff-irrelevant, so that $M$ is indeed a best response to $L$.

To show that the $(\sigma, \tau)$-constrained efficient strategy profile $(L, M)$ is $(\sigma, \tau)$-deferential, we must prove that periods $\sigma$ and $\tau$ are reached with positive probability. Assume, by way of contradiction, that one constraint does not bind, e.g., the conversation ends in the $\mathcal{A}$-subgame at some earlier (say, odd) period $t < \tau$; that is, all remaining types of Lones concede in period $t$. We now define an alternative $(\sigma, \tau)$-stopping profile $(L', M')$ with higher ex-ante payoffs. Consider Moritz' strategy $M'$ that coincides with $M$ up to period $t-1$ but for which all types of Moritz concede in period $t+1$. As period $t+1$ is off path, $(L, M')$ is also constrained efficient. Since Moritz is about to concede, strong enough types of Lones, who are quite convinced of guilt, are then willing to wait one more period to secure a conviction;[22] thus, for any best-response $L'$ to $M'$, strategy profile $(L', M')$ is a $(\sigma, \tau)$-stopping profile yielding higher ex-ante payoffs than $(L, M)$. This contradicts the constrained efficiency of $(L, M)$. An analogue argument shows that the unconstrained efficient strategy profile must be a communicative equilibrium.

## B.5 Stability: Proof of Theorem 4

More generally, consider a $(\sigma, \tau)$-deferential equilibrium with arbitrary off-path cutoffs $x_{\tau-1} \leq x_{\tau+1} \leq x_{\tau+3} \leq \cdots$. Let $\tilde{\ell}$ be the weakest type of Lones who would hold out in period $\tau$ for some alternative off-path behavior of Moritz. By (5), illustrated

[21] By Lemma 3, $L'$ is also agreeable.

[22] More formally, Lones' propensity to hold out (5) with $\bar{x} = \infty$ is positive for sufficiently high $y$ (for then $\pi(y, x)$ is near 1).

in Figure 2, type $\tilde{\ell}$'s propensity to hold out is maximized by $x_{\tau+1} = \tilde{\ell} + k$, and so $P^{\mathcal{A}}(x_{\tau-1}, \tilde{\ell}, \tilde{\ell} + k) = 0$.

Assume counterfactually that this equilbrium satisfies forward induction. Then Moritz' beliefs in period $\tau + 1$ are restricted to $[\tilde{\ell}, \infty)$; thus, types $m < \tilde{\ell} + k$ optimally concede, and so $\delta \equiv x_{\tau+1} - (\tilde{\ell} + k) > 0$. This leads to the contradiction that Lones' type $x_{\tau+1} - k$ strictly prefers to hold out in period $\tau$ because

$$0 = P^{\mathcal{A}}(x_{\tau-1}, \tilde{\ell}, \tilde{\ell} + k) < P^{\mathcal{A}}(x_{\tau-1} + \delta, x_{\tau+1} - k, x_{\tau+1}) < P^{\mathcal{A}}(x_{\tau-1}, x_{\tau+1} - k, x_{\tau+1}),$$

where the first inequality follows from Lemma 4(b), and the second inequality follows from Lemma 4(a).

## B.6 Uniqueness: Proof of Theorem 5

Recall the definition of the *propensity to hold out*:[23]

$$P^{\mathcal{A}}(\underline{x}, y, \bar{x}) = \int_{\underline{x}}^{\bar{x}} (2\pi(y, x) - 1 - \kappa) f(x|y) dx - \int_{\bar{x}}^{\infty} 2\kappa f(x|y) dx \tag{14}$$

**Lemma 4 (Propensity Function)**
*(a) When $\underline{x} + k < y < \bar{x} - k$, $P^{\mathcal{A}}$ is strictly decreasing and differentiable in $\underline{x}$, $\bar{x}$.*
*(b) If $P^{\mathcal{A}}(\underline{x}, y, \bar{x}) = 0$ then $P^{\mathcal{A}}(\underline{x} + \delta, y + \delta, \bar{x} + \delta) > 0$ for any $\delta > 0$.*

**Proof** These properties are intuitive in light of Figure 2. More formally, the partial derivatives of $P^{\mathcal{A}}$ with respect to $\underline{x}$ and $\bar{x}$ are given by:

$$P^{\mathcal{A}}_{\underline{x}} = -(2\pi(y, \underline{x}) - 1 - \kappa) f(\underline{x}|y) \quad \text{and} \quad P^{\mathcal{A}}_{\bar{x}} = (2\pi(y, \bar{x}) - 1 - \kappa + 2\kappa) f(\bar{x}|y) \tag{15}$$

Given $k = \log \frac{1+\kappa}{1-\kappa}$ and $2\pi(y, x) - 1 = \frac{e^{y-x}-1}{e^{y-x}+1}$, the assumption $\underline{x} + k < y < \bar{x} - k$ implies that these expressions are both negative. Intuitively, holding out and reaching a convict verdict against the opponent's types $x \in [\underline{x}, \bar{x}]$ is less attractive against stronger types.

Part (b): define the function $\psi$ by $\psi(x) := 2\pi(y, x) - 1 - \kappa$ for $x \in [\underline{x}, \overline{x}]$ and $\psi(x) := -2\kappa$ for $x \in [\bar{x}, \infty)$, so that

$$P^{\mathcal{A}}(\underline{x} + \delta, y + \delta, \bar{x} + \delta) = \int_{\underline{x}}^{\infty} \psi(x) f(x + \delta | y + \delta) dx$$

By Lemma 1, the density $f(x + \delta|y + \delta)$ is log-submodular in $x$ and $\delta$, i.e., greater $\delta$ shifts the density $f$ down in the MLRP sense; the integrand $\psi(x)$ single-crosses from

[23] We also think of $P$ as a function of $\kappa$, but we shall suppress this argument for now.

above in $\delta$, and so $P^{\mathcal{A}}(\underline{x}+\delta, y+\delta, \bar{x}+\delta)$ single-crosses from below by Karlin and Rubin (1956), Lemma 1. $\square$

Analogously, the propensity to hold out in the $\mathcal{C}$-subgame when the opponent's type is below $\bar{x}$ and his types $x \in [\underline{x}, \bar{x}]$ are about to concede, $P^{\mathcal{C}}(\underline{x}, y, \bar{x})$ as defined in (6), strictly increases in $\bar{x}$ and $\underline{x}$ and diagonally single-crosses from above. Finally, Moritz's propensity to initially vote $\mathcal{A}$, $P^0(\underline{x}, y, \bar{x})$ as defined in (8), strictly increases in Lones' first cutoff types $\underline{x}, \bar{x}$ and diagonally single-crosses from below.

By Theorem 1, cutoff vectors $X$ are an equilibrium iff

$$P^{\mathcal{A}}(x_{t-1}, x_t, x_{t+1}) = 0, \qquad P^0(x_{-1}, x_0, x_1) = 0, \qquad P^{\mathcal{C}}(x_{-s-1}, x_{-s}, x_{-s+1}) = 0 \quad (16)$$

for all $s, t$ with $-\infty < x_{-s} < x_t < \infty$. By (10) thresholds are separated: $x_{t+1} - x_t > k$ for all $t$. By the monotonicity of the propensity functions, any two adjacent cutoffs recursively determine all others. We now explore these "shooting functions".

Specifically, define $\mathcal{D} \subset \mathbb{R}^2$ as the set of pairs $(x, y)$ with $x < y - k$ for which $P^{\mathcal{A}}(x, y, \bar{x}) = 0$ has a solution $\bar{x} \in [y + k, \infty]$. By the strict monotonicity of $P^{\mathcal{A}}$ in $\bar{x}$, such $\bar{x}$ is unique and we can define the differentiable *shooting function* $\phi$ on domain $\mathcal{D}$:

$$P^{\mathcal{A}}(x, y, \phi(x, y)) = 0. \quad (17)$$

So $(x_{t-1}, x_t)$ "shoots" to $x_{t+1} = \phi(x_{t-1}, x_t)$, for $t = 1, 2, 3, \ldots$.

In light of (16), we wish to anchor our recursion at $(x_0, x_1)$; therefore, we shall define an initial period shooting function $\phi^0(y, x)$ by $P^0(\phi^0(y, x), y, x) = 0$.

Finally, observe that in the $\mathcal{C}$-subgame, $P^{\mathcal{C}}(-\phi(-x, -y), -y, -x) = 0$, by symmetry of the density $f$. So as long as $-\infty < x_{-s} < x_t < \infty$, equilibrium cutoffs obey:

$$x_{t+1} = \phi(x_{t-1}, x_t), \qquad x_{-1} = \phi^0(x_0, x_1), \qquad x_{-(s+1)} = -\phi(-x_{-(s-1)}, -x_{-s}). \quad (18)$$

Note that $x_\tau = \infty$ and $x_{-\sigma} = -\infty$ for the $(\sigma, \tau)$-deferential equilibrium with finite $\sigma, \tau$, while all cut-offs are finite in a communicative equilibrium.

Let $a(y) \equiv \inf\{x : (x, y) \in \mathcal{D}\}$ denote the left boundary of $\mathcal{D}$.

**Lemma 5** *For large $y$, both $a(y) \in (-\infty, y - k)$ and $y - a(y)$ decrease.*

*Proof*: Consider the propensity to hold out when all opponents' types concede in the next period, $P^{\mathcal{A}}(-\infty, y, \infty)$. By Lemma 2, this function single-crosses in $y$, and by the definition (14) of $P^{\mathcal{A}}$, it is positive for large $y$ because the integrand $2\pi(y, x) - 1 - \kappa \gtrless$

0 for all $x \lneqq y - k$ (see Figure 2). Fix $y$ with $P^{\mathcal{A}}(-\infty, y, \infty) > 0$. Next, since $2\pi(y, x) - 1 - \kappa < 0$ for all $x > y - k$, we have $P^{\mathcal{A}}(y - k, y, \infty) < 0$. Finally, $P^{\mathcal{A}}(\underline{x}, y, \infty)$ is continuous in $\underline{x}$ on $[-\infty, y - k]$ by Lemma 4$(a)$. We conclude from the Intermediate Value Theorem that there exists $x^* \in (-\infty, y-k)$ with $P^{\mathcal{A}}(x^*, y, \infty) = 0$. In fact, since $P^{\mathcal{A}}$ decreases in its third argument by Lemma 4(a), $x^*$ is the least root $x$ for which $P^{\mathcal{A}}(x, y, \bar{x}) = 0$ has a solution $\bar{x} \in [y + k, \infty]$. So $a(y) = x^*$.

Since $P^{\mathcal{A}}(a(y), y, \infty) = 0$, we have $P^{\mathcal{A}}(a(y) + \delta, y + \delta, \infty) > 0$ for all $\delta > 0$ by Lemma 4(b). Thus, $P^{\mathcal{A}}(\underline{x}, y + \delta, \bar{x}) > 0$ for all $\underline{x} \leq a(y) + \delta$ and $\bar{x} \in [y + k, \infty]$. So $a(y + \delta) > a(y) + \delta$. Thus, $y - a(y)$ is decreasing. □

For any $\delta > k$ and $\epsilon \geq 0$, we let $\mathcal{D}_{\delta,\epsilon}$ denote a subdomain of $\mathcal{D}$ in $\mathbb{R}^2_+$, bounded away from the diagonal, where each coordinate gap of (17) is at least $\delta$, namely $y - x \geq \delta$ and $\phi(x, y) - y \geq \delta$, but where the second gap is not more than $\epsilon$ bigger, namely, $\phi(x, y) - y \leq y - x + \epsilon$.

We next show that $\phi$ strictly decreases in the horizontal $x$ direction, and increases along the diagonal at more than unit speed.

**Lemma 6 (Properties of the Shooting Function)**
*(a) On the domain $\mathcal{D}$, we have $\phi_x(x, y) < 0$.*
*(b) More strongly, for any $\delta > k$, there is $\epsilon > 0$, such that $\phi_x(x, y) \leq -(1 + \epsilon)$ on the subdomain $\mathcal{D}_{\delta,\epsilon}$. Also, there exists $\epsilon' > 0$ such that for every $\delta' > k$, we have $\phi_x(x, y) \leq -(1 + \epsilon')$ on the subdomain $\mathcal{D}_{\delta',0}$.*
*(c) On the domain $\mathcal{D}$, we have $\phi(x + \delta, y + \delta) > \phi(x, y) + \delta$ for any $\delta > 0$.*

*Proof of* $(a)$: Write $\bar{x} = \phi(x, y)$ and $\underline{x} = x$, and apply the implicit function theorem to (17), substituting (15) for the partial derivatives of $P^{\mathcal{A}}$. Then

$$-\phi_x(x, y) = \frac{P^{\mathcal{A}}_{\underline{x}}(\underline{x}, y, \bar{x})}{P^{\mathcal{A}}_{\bar{x}}(\underline{x}, y, \bar{x})} = \frac{2\pi(y, \underline{x}) - 1 - \kappa}{1 - 2\pi(y, \bar{x}) - \kappa} \cdot \frac{f(\underline{x}|y)}{f(\bar{x}|y)}. \tag{19}$$

This ratio is positive on $\mathcal{D}$ since both the numerator and denominator are positive.

*Proof of* $(b)$: More strongly, given (1), we see that $2\pi(y, x) - 1 = \frac{e^{y-x}-1}{e^{y-x}+1} = 1 - 2\pi(x, y)$. So for any $(\underline{x}, y) \in \mathcal{D}_{\delta,0}$, since $y - \underline{x} \geq \bar{x} - y \geq \delta > k = \log(1 + \kappa)/(1 - \kappa)$, we have

$$2\pi(y, \underline{x}) - 1 - \kappa \geq 1 - 2\pi(y, \bar{x}) - \kappa \geq \frac{e^\delta - 1}{e^\delta + 1} - \kappa > 0$$

Moreover, $f(x|y) = f(x)r(x, y)$ is decreasing in $x$ when $x, y > 0$, since $\partial_x r(x, y) = \frac{2}{1+e^y} \frac{e^x(1-e^y)}{(1+e^x)^2} < 0$, and $f$ is log-concave and symmetric and thus continuously decreasing.

Thus, (19) strictly exceeds 1 for $(\underline{x}, y) \in \mathcal{D}_{\delta,0}$. By continuity of (19), there exists $\epsilon > 0$ such that (19) strictly exceeds $1 + \epsilon$ for $(\underline{x}, y) \in \mathcal{D}_{\delta,\epsilon}$.

*Proof of* (*c*): First, $P^{\mathcal{A}}(\underline{x}+\delta, y+\delta, \phi(\underline{x}, y)+\delta) > P^{\mathcal{A}}(\underline{x}, y, \phi(\underline{x}, y)) = 0$ by Lemma 4(*b*). Then $\phi(\underline{x} + \delta, y + \delta) > \phi(\underline{x}, y) + \delta$ by Lemma 4(*a*). □

Analogous arguments to Lemma 6(a) and (c) show that $\phi^0(y, x)$ strictly decreases in $x$ and increases diagonally at more than unit speed, $\phi^0(y + \delta, x + \delta) > \phi^0(y, x) + \delta$.

**Lemma 7** *In a communicative equilibrium* (18)*, $x_{t+1} - x_t$ strictly decreases in $t > 0$.*

**Proof** Put $\delta_t \equiv x_{t+1} - x_t$ and note that $\delta_t > k$ by (10). Now assume, to the contrary, a communicative equilibrium $(x_t)$ with $\Delta_t \equiv \delta_{t+1} - \delta_t \geq 0$ for some period $t$. Then

$$\begin{aligned} \delta_{t+2} &= x_{t+3} - x_{t+2} = \phi(x_{t+1}, x_{t+2}) - x_{t+2} \geq \phi(x_{t+1} + \Delta_t, x_{t+2}) - x_{t+2} \\ &= \phi(x_t + \delta_{t+1}, x_{t+1} + \delta_{t+1}) - x_{t+2} > \phi(x_t, x_{t+1}) + \delta_{t+1} - x_{t+2} = \delta_{t+1} \end{aligned} \tag{20}$$

by Lemma 6(a) and (c). So $\Delta_{t+1} > 0$. By induction, $\Delta_s > 0$ for all $s > t$.

We now argue more strongly that $\Delta_s$ is bounded away from zero. Since $\delta_s$ is positive and increasing, there exists $\underline{t}$ such that $x_s > 0$ for all $s > t'$ and, obviously, $\delta_s > \delta_{\underline{t}} \equiv \delta$. By Lemma 6(b) there exists $\epsilon > 0$ such that

$$\phi(x_{s+1}, x_{s+2}) > \phi(x_{s+1} + \Delta_s, x_{s+2}) + (1 + \epsilon)\Delta_t.$$

if $(x_{s+1}, x_{s+2}) \in \mathcal{D}_{\delta,\epsilon}$. Substituting this inequality for the first inequality of (20), we conclude more strongly that $\Delta_{s+1} \geq (1 + \epsilon)\Delta_s$. Otherwise, if $(x_{s+1}, x_{s+2}) \notin \mathcal{D}_{\delta,\epsilon}$, then the definition of $\mathcal{D}_{\delta,\epsilon}$ together with $x_s > 0$ and $\delta_{s+1}, \delta_{s+2} > \delta$ implies $\delta_{s+2} > \delta_{s+1} + \epsilon$. That is, $\Delta_{s+1} > \epsilon$. Altogether, we have $\Delta_{s+1} \geq \min\{\epsilon, (1 + \epsilon)\Delta_s\}$.

Thus, the step size $\delta_s \geq \delta_{\underline{t}} + (s - \underline{t}) \cdot \min\{\epsilon, \Delta_{\underline{t}}\}$ diverges. By Lemma 5, the width $y - a(y)$ of the domain $\mathcal{D}$ is falling, and so bounded for large $y$. So $x_{s+1} - x_s = \delta_s > x_{s+1} - a(x_{s+1})$ for large $s$. Then $x_s < a(x_{s+1})$, and so $(x_s, x_{s+1}) \notin \mathcal{D}$. So $x_{s+2} = \phi(x_s, x_{s+1})$ is not well-defined, so $(x_t)$ is not a communicative equilibrium. □

Lemma 7 implies that any two consecutive, positive cutoffs in a communicative equilibrium satisfy $(x_t, x_{t+1}) \in \mathcal{D}_{\delta_t,0}$, so that the second part of Lemma 6(b) applies to $\phi(x_t, x_{t+1})$.

**Lemma 8 (Uniqueness)** *There is a unique communicative equilibrium. Also, there is a unique $(\sigma, \tau)$-deferential equilibrium for any $(\sigma, \tau)$.*

**Proof** Assume two communicative equilibria $(x'_t) \neq (x_t)$, with $x'_0 \geq x_0$. First assume $\Delta_0 \equiv (x'_1 - x_1) - (x'_0 - x_0) \geq 0$, with strict inequality if $x'_0 = x_0$. As in the derivation of (20),

$$x'_2 = \phi(x'_0, x'_1) \geq \phi(x'_0 + \Delta_0, x'_1) = \phi(x_0 + (x'_1 - x_1), x_1 + (x'_1 - x_1)) > x_2 + (x'_1 - x_1) \quad (21)$$

by Lemma 6(a) and (c). By induction, we conclude $\Delta_t \equiv (x'_{t+1} - x_{t+1}) - (x'_t - x_t) > 0$.

We now argue that $\Delta_t \to \infty$. As $\delta'_t \equiv x'_{t+1} - x'_t > k$ by (10), there exists $\underline{t}$ such that $x_t > 0$ for all $t > \underline{t}$. By the second part of Lemma 6(b), there exists $\epsilon' > 0$ such that

$$x'_{t+2} = \phi(x'_t, x'_{t+1}) \geq \phi(x'_t + \Delta_t, x'_{t+1}) + (1 + \epsilon')\Delta_t.$$

Proceeding as in (21) implies $\Delta_{t+1} > (1 + \epsilon')\Delta_t$ for all $t > \underline{t}$ and thus $\Delta_t \to \infty$.

Then $\delta_t \equiv x_{t+1} - x_t = \delta'_t - \Delta_t < 0$ for large $t$ because $\Delta_t \to \infty$ while $\delta'_t$ decreases by Lemma 7. Thus, $(x_t)$ violates (10) and cannot be an equilibrium.

Now consider the alternative case $\Delta_0 \equiv (x'_1 - x_1) - (x'_0 - x_0) < 0$. Since $\phi^0(y, x)$ strictly decreases in $x$ and increases diagonally at more than unit speed, $\phi^0(y+\delta, x+\delta) > \phi^0(y, x) + \delta$, as in the derivation of (20) and (21):

$$x'_{-1} = \phi^0(x'_0, x'_1) > \phi^0(x'_0, x'_1 - \Delta_0) = \phi^0(x_0 + (x'_0 - x_0), x_1 + (x'_0 - x_0)) > x_{-1} + (x'_0 - x_0).$$

So $\Delta_{-1} \equiv (x'_{-1} - x_{-1}) - (x'_0 - x_0) > 0$ and analogous arguments as above show that $\Delta_{-t} \to \infty$ diverges. Again $(x'_t)$ and $(x_t)$ cannot both be communicative equilibria.

Now assume two $(\sigma, \tau)$-deferential equilibria $(x'_t) \neq (x_t)$, with $x'_0 \geq x_0$ with $\Delta_0 \equiv (x'_1 - x_1) - (x'_0 - x_0) \geq 0$,[24] with strict inequality if $x'_0 = x_0$. As for the communicative equilibrium we get $\Delta_t \equiv (x'_{t+1} - x_{t+1}) - (x'_t - x_t) > 0$.

As $(x_t)$ is a $(\sigma, \tau)$-deferential equilibrium, we have $P^{\mathcal{A}}(x_{\tau-2}, x_{\tau-1}, \infty) = 0$, so $x_{\tau-2} = a(x_{\tau-1})$. As the slope of $a(y)$ exceeds one (by Lemma 5) and $\Delta_{\tau-2} > 0$ we have

$$a(x'_{\tau-1}) = a(x_{\tau-1} + (x'_{\tau-1} - x_{\tau-1})) > a(x_{\tau-1}) + (x'_{\tau-1} - x_{\tau-1}) > x_{\tau-2} + (x'_{\tau-2} - x_{\tau-2}) = x'_{\tau-2}.$$

Thus, $P^{\mathcal{A}}(x'_{\tau-2}, x'_{\tau-1}, \infty) > 0$ so $(x'_t)$ is not a $(\sigma, \tau)$-deferential equilibrium. □

[24] As with the communicative equilibrium, if $\Delta_0 < 0$ we can derive an analogous contradiction in the $\mathcal{C}$-subgame.

## B.7 Comparative Static: Proof of Proposition 4

Write the propensity to hold out and the shooting function also as function of waiting costs $\kappa$, $P^{\mathcal{A}}(x,y,\bar{x}|\kappa)$, $\phi(x,y|\kappa)$. Since $P^{\mathcal{A}}$ falls in $\kappa$ by (14), and falls in $\bar{x}$ by Lemma 4(a), the implicit function theorem implies

$$\phi_\kappa(x,y|\kappa) = -\frac{P^{\mathcal{A}}_\kappa(x,y,\bar{x}|\kappa)}{P^{\mathcal{A}}_{\bar{x}}(x,y,\bar{x}|\kappa)} < 0.$$

Now fix costs $\kappa' < \kappa$ and consider the unique communicative equilibria $(x'_t), (x_t)$. By symmetry we have $x'_0 = x_0 = 0$. We need to show $x'_t < x_t$ for all $t > 0$. Assume to the contrary that there exists $t > 0$ with $x'_t \geq x_t$ and choose $t$ minimal, so that $x'_{t-1} \leq x_{t-1}$. Let $\Delta_{t-1} \equiv (x'_t - x_t) - (x'_{t-1} - x_{t-1})$. Then, $\phi(x,y|\kappa') > \phi(x,y|\kappa)$ and the second part of Lemma 6(b) imply that there exists $\epsilon' > 0$ with

$$x'_{t+1} \;=\; \phi(x'_{t-1}, x'_t|\kappa') > \phi(x'_{t-1}, x'_t|\kappa) > \phi(x'_{t-1} + \Delta_{t-1}, x'_t|\kappa) + (1+\epsilon')\Delta_{t-1}$$

Proceeding as in (21) implies $\Delta_t > (1+\epsilon')\Delta_{t-1}$ and thus inductively $\Delta_t \to \infty$, leading to the same contradiction as in the proof of Lemma 8. This contradiction implies that the cutoffs in the communicative equilibrium satisfy $x'_t < x_t$ for all $t > 0$. The proof for the symmetric, $(\tau,\tau)$-deferential equilibrium is similar to that of Lemma 8.

## B.8 Fast Conversations: Proof of Proposition 5

We first establish two auxiliary results. First, $|2\pi(\ell,m) - 1| = \frac{|e^{\ell-m}-1|}{e^{\ell-m}+1}$ by (1), which vanishes when $\ell = m$. Applying $\frac{d}{dx}\frac{e^x-1}{e^x+1} = \frac{2e^x}{(e^x+1)^2} = \frac{1}{2}$ at $x = 0$ yields the approximation:

**Claim 2** $|2\pi(\ell,m) - 1| = \frac{1}{2}|\ell - m| + O(|\ell - m|^2)$.

Next, since $\frac{\partial}{\partial \ell}\frac{2(e^\ell+e^m)}{(1+e^\ell)(1+e^m)} = \frac{2e^\ell(1-e^m)}{(1+e^\ell)^2(1+e^m)}$, and similarly for $m$, we have:

**Claim 3** *For $\ell, m \geq 0$, $r(\ell,m)$ decreases, and obeys $r(\ell,\ell) = \frac{4e^\ell}{(1+e^\ell)^2} \leq 4e^{-\ell}$.*

A. An Asymptotic Upper Bound for $c(\kappa)$. Define $(x_t)$ by $x_t = t\kappa^{1/3}$. We define $\bar{\theta}$ so that the associated total costs are less than $\bar{\theta}\kappa^{2/3}$.

Decision costs arise in (the odd) period $t+1$ of the $\mathcal{A}$-subgame if Lones wrongfully concedes to acquit a defendant who should be convicted, i.e., $m < \ell$. Ex ante, these are

$$\int_{x_t < m < \ell < x_{t+1}} (2\pi(\ell,m) - 1) r(\ell,m) f(\ell) f(m) d\ell dm. \tag{22}$$

By Claim 2, the loss is at most $\ell - m < x_{t+1} - x_t = \kappa^{1/3}$ for small $\kappa$. By Claim 3, the correlation is at most $4e^{-t\kappa^{1/3}}$. The density is at most $\bar{f} = \max_\ell f(\ell)$. Since the Lebesgue measure of the triangular domain of (22) equals $\kappa^{2/3}/2$, losses (22) are bounded above by $(\kappa^{2/3}/2)\kappa^{1/3}4e^{-t\kappa^{1/3}}\bar{f}^2 = 2\kappa\bar{f}^2(e^{-\kappa^{1/3}})^t$.

The decision costs are similarly bounded in (the even) period $t+2$ when Moritz wrongfully concedes to convict the defendant, and in odd and even periods of the $\mathcal{C}$-subgame. For small $\kappa$, given $(1-e^{-x}) \geq x/2$ for small $x$, total decision costs are at most

$$4\kappa\bar{f}^2 \sum_{t=0}^{\infty} (e^{-\kappa^{1/3}})^t = \frac{4\kappa\bar{f}^2}{1-e^{-\kappa^{1/3}}} \leq 8\bar{f}^2\kappa^{2/3}$$

The probability of reaching period $t+1$ (in either subgame) is given by

$$\int_{m,\ell>t\kappa^{1/3}} r(\ell,m)f(\ell)f(m)d\ell dm \leq r(t\kappa^{1/3}, t\kappa^{1/3}) \int f(\ell)d\ell \int f(m)dm \leq 4e^{-t\kappa^{1/3}}$$

since $r(\ell,m)$ decreases, and $f \geq 0$. So waiting costs are bounded above by

$$\sum_{t=0}^{\infty} 8(e^{-\kappa^{1/3}})^t\kappa \leq \frac{8}{1-e^{-\kappa^{1/3}}}\kappa \leq 16\kappa^{2/3}.$$

Thus, $c(\kappa) \leq \bar{\theta}\kappa^{2/3}$ for $\bar{\theta} \equiv 8\bar{f}^2 + 16$.

B. An Asymptotic Lower Bound for $c(\kappa)$. Fix any threshold $\omega > 0$. Let $\underline{h} = \min\{r(\ell,m)f(\ell)f(m) : \ell, m \in [0,\omega]\} > 0$ be the lower bound on the joint density for "low types" $[0,\omega]$, and $\underline{p} = P(\ell, m > \omega) > 0$ the lower bound on the chance of "high types" $[\omega,\infty)$. We deduce a lower bound $\underline{\theta}\kappa^{2/3}$ across all strategy profiles $(x_t)$ for the sum of decision costs of low types and waiting costs of high types.

By Claim 2, realized decision costs are greater than $(\ell-m)/4$ when $\ell - m$ is small. Thus, expected decision costs in (odd) period $t+1$, as given by (22), are bounded below by

$$\underline{h}\int_{x_t<m<\ell<x_{t+1}} \frac{\ell-m}{4} d\ell dm = \frac{\underline{h}(x_{t+1}-x_t)^3}{24}.$$

when $x_t, x_{t+1} \in [0,\omega]$, and $x_{t+1} - x_t$ is small enough. Fix a strategy profile $(x_t)$ and assume WLOG $x_0 \leq 0$. Define $\underline{T} = \max\{t : x_t \leq 0)$ and $\bar{T} = \min\{t : x_t \geq \omega\}$. Expected decision costs of low types are at least

$$\frac{\underline{h}}{24}\left[(x_{\underline{T}+1} - 0)^3 + \sum_{t=\underline{T}+1}^{\bar{T}-2} (x_{t+1}-x_t)^3 + (\omega - x_{\bar{T}-1})^3\right].$$

Put $T = \bar{T} - \underline{T}$. This sum of $T$ cubic terms is minimized when all terms equal $(\omega/T)^3$. So expected decision costs for low types are bounded below by

$$\frac{\underline{h}}{24} T \left(\frac{\omega}{T}\right)^3 \equiv \theta_1 T^{-2}.$$

Since high types wait at least $T$ periods until agreement, the associated waiting costs exceed $\underline{p} T \kappa$. The lower bound on total costs, $\theta_1 T^{-2} + \underline{p} T \kappa$, is minimized by $T^* = (2\theta_1/(\underline{p}\kappa))^{1/3}$. So $c(\kappa) \geq \underline{\theta}\kappa^{2/3}$ where $\underline{\theta} \equiv 3\theta_1(\underline{p}\kappa/(2\theta_1))^{2/3}$.

# References


Albrecht, J., A. Anderson, and S. Vroman (2010): "Search by Committee," *Journal of Economic Theory*, 145, 1386–1407.

Aumann, R., and S. Hart (2003): "Long Cheap Talk," *Econometrica*, 71, 1619–1660.

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

Austen-Smith, D., and T. Feddersen (2006): "Deliberation, Preference Uncertainty, and Voting Rules," *American Political Science Review*, 100, 209–217.

Cho, I.-K. (1987): "A Refinement of Sequential Equilibrium," *Econometrica*, 55, 1367–1389.

Compte, O., and P. Jehiel (2010): "Bargaining and Majority Rules: A Collective Search Perspective," *Journal of Political Economy*, 118, 189–221.

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 H. Haller (1990): "Learning How To Cooperate: Optimal Play in Repeated Coordination Games," *Econometrica*, 58, 571–595.

Crawford, V. P., and J. Sobel (1982): "Strategic Information Transmission," *Econometrica*, 50, 1431–1451.

Damiano, E., H. Li, and W. Suen (2012): "Optimal Deadlines for Agreements," *Theoretical Economics*, 7, 357–393.

Feddersen, T., and W. Pesendorfer (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.

——— (1998): "The Inferiority of Unanimous Jury Verdicts under Strategic Voting," *American Political Science Review*, 92, 23–35.

Gerardi, D., and L. Yariv (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 B. Szentes (2009): "Optimal voting schemes with costly information acquisition," *Journal of Economic Theory*, 144, 36–68.

Gul, F., and R. Lundholm (1995): "Endogenous Timing and the Clustering of Agents Decisions," *Journal of Political Economy*, 103, 1039–1066.

Gul, F., and W. Pesendorfer (2012): "The War of Information," *Review of Economic Studies*, 79, 707–734.

Karlin, S., and B. Rubin (1956): "The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio," *Annals of Mathematical Statistics*, 27, 272–299.

Li, H., S. Rosen, and S. Wing (2001): "Conflicts and Common Interests in Committees," *American Economic Review*, 91, 1478–1497.

Li, H., and S. Wing (2009): "Decision-making in Committees," *Canadian Journal of Economics*, 42, 359–392.

Lizzeri, A., and L. Yariv (2012): "Sequential Deliberation," mimeo.

Moldovanu, B., and X. Shi (2013): "Specialization and partisanship in committee search," *Theoretical Economics*, 8, 751–774.

Moscarini, G., and L. Smith (2002): "The Law of Large Demand for Information," *Econometrica*, 70, 2351–2366.

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.

Smith, L., P. Sorensen, and J. Tian (2012): "Informational Herding, Optimal Experimentation, and Contrarianism," mimeo.

Strulovici, B. (2010): "Learning while Voting: Determinants of Collective Experimentation," *Econometrica*, 78, 933–971.

Wald, A. (1947): *Sequential Analysis*. John Wiley and Sons, New York, 1st edn.

Weitzman, M. (1979): "Optimal Search for the Best Alternative," *Econometrica*, 47, 641–654.