Excessive Harm

Steffen Huck*

July 29, 2026

The horror! The horror!
Joseph Conrad (1899)

Denk ich an Deutschland in der Nacht,
Dann bin ich um den Schlaf gebracht.
Heinrich Heine (1844)

Abstract

A father punishes a son who has done a misdeed; a mother can sanction the father to protect him. Both sides harbour private information: the father may be just or harsh, the mother's resolve to carry out sanctions strong or weak. I characterise when protection fails. Any protection requires early sanctions. Yet, sanctioning can fail in two ways: through silence, when sanctions cost more than they spare, and through unmasking, when the harsh father screens the mothers and only the resolute mother can protect her son. The model speaks to international law, where punisher and protector are states.

1 Introduction

A son has conducted a misdeed. A father wants to punish—perhaps justly, perhaps far beyond. A mother stands by watching how punishments unfold. She has one instrument though: she can sanction the father after each round of punishments.

I shall stick to this allegory throughout, although the applications I am interested in are much broader, including an application to international law where a state

*WZB Berlin Social Science Center and UCL. I thank Tilman Fries for exceptionally perceptive comments on an earlier draft.

actor is represented by the father, a population that bears the punishment by the son, and third states by the mother.

The father has two types—a just father who wants to punish justly and a harsh father who wants to punish harshly. The mother also has two types, which vary in the costs of sanctioning: a strong mother with lower costs and a weak mother with higher costs. The mother's main objective is to protect her son from excessive punishment, and the paper's main objective is to characterise the situations where she fails in this endeavour.

There is a trivial failure and a not-so-trivial one. The trivial failure arises when the harsh father's zeal is such that even a strong mother cannot deter him fully. The non-trivial failure arises through an interplay of the two sources of incomplete information. There are parameter bands where a strong mother can shelter her son but a weak mother cannot—because the pooling equilibria that would protect all sons break down. The father's zeal will play a role here, too.

This paper was written with Gaza in view, and it would be coy to pretend otherwise. But the reader should know exactly what is, and is not, being claimed. Father, mother, and son are mnemonic labels, chosen for compression, but they apply to any similar situation in which there is a punisher, a passive entity to be punished, and a third party that can intervene through sanctions. The model does not adjudicate the conflict. It takes as given that some level of punishment is just and that the harsh punisher's ideal exceeds it. It deliberately gives the son no moves—an abstraction isolating the punishment–sanction relation.

How time unfolds is an important determinant of that relation. In order to simplify as much as possible without losing the gist, I study a model where the father can punish his son in two periods. After the first punishment the mother can immediately impose a sanction on the father and she can announce a sanctioning scheme for excessive punishments in round 2. Whether these later sanctions are carried out depends on the mother's sanctioning costs.

This two-period structure allows for the analysis of the interplay between the mother's uncertainty about the father's type and the father's uncertainty about the mother's type. Weak mothers will have an incentive to pool with strong mothers, but harsh fathers may, as it turns out, prefer to screen the mothers and reveal their type by doing so. If that screening succeeds—because the weak mother cannot afford to pool with the strong mother—the non-trivial path to

excessive punishment arises.

The paper proceeds as follows. Section 2 places the model in the literature. Section 3 sets out the model, and Section 4 solves the game through backward induction from the second stage onwards, that is, after the father's first strike; it establishes a first key proposition illuminating the trivial and the non-trivial case. Section 5 then examines the father's first strike and states the main theorem, characterising the situations where excessive punishments unfold. Section 6 discusses the results and Section 7 offers a coda.

2 Related literature

This section maps the neighbourhoods the paper borders. It has four main ones.

Signalling, and the price of being believed. The paper's machinery descends from Spence (1973): a costly action separates types when its cost differs across them. Riley (1979) characterised the least-cost separating outcome our equilibrium selects, and the refinement that disciplines beliefs off the path is the Intuitive Criterion of Cho and Kreps (1987), with Banks and Sobel (1987) the neighbouring divinity refinement. Two features place the present model in a particular corner of that tradition. First, the period-one sanction is, in the language of the literature, *burned money*: an action whose only equilibrium function is to be seen to have been costly. Ben-Porath and Dekel (1992) showed how the ability to burn money before play selects outcomes; Austen-Smith and Banks (2000) married burning to cheap talk and established when the match adds credibility. The mother's initial sanction is just that—burned money. The mother's announcement of a sanctioning scheme for further potentially excessive punishments could be modeled as cheap talk in the tradition of Crawford and Sobel (1982) but we will depart from that by assuming posturing costs in the spirit of the costly-talk models of Kartik et al. (2007) and Kartik (2009).

Resolve, and how to demonstrate it. That threats are instruments, that their credibility is the question, and that commitment is a strategic act are the inheritance of Schelling (1960, 1966)—who, it is worth recalling, reached repeatedly for the disciplining of children as his illustration of choice: the family frame has a pedigree. The formal study of resolve descends through the reputation models of Kreps and Wilson (1982) and Milgrom and Roberts (1982), where a type that would carry threats out lends its shadow to types that would not.

Fearon (1995) established war among rational actors as an equilibrium phenomenon; Powell (2006) located its deepest cause in commitment problems. Closest to our instrument pair is Fearon (1997), whose distinction between *sinking costs* and *tying hands* maps almost exactly onto our two types of sanctions, the immediate one and the announced one; the period-one sanction is sunk, the announced rate ties no hands at all—it is the limiting case Fearon's dichotomy leaves implicit, speech whose only collateral is the posturing paid to utter it. Audience-cost and reputational theories of communication (Fearon, 1994; Schultz, 1998; Sartori, 2002) supply what our announcements conspicuously lack, an external penalty for bluffing; the model shows how much sorting survives on posturing costs alone. Slantchev (2005) studies military instruments that signal because they hurt, Sechser (2010) why strength itself can spoil coercion, Kydd (2005) how trust is built from costly gestures, Abreu and Gul (2000) how obstinate types shape bargaining, and Baliga and Sjöström (2004) how cheap talk tempers arms races. We add to this family a third party: resolve here is demonstrated not to deter an opponent from attacking oneself, but to protect someone else.

Sanctions. The economics of sanctions has a long tradition with Kaempfer and Lowenberg (1988) and Eaton and Engers (1992) being a good starting point, the latter the canonical model of a sender extracting compliance through threatened harm, compressed in Eaton and Engers (1999). Tsebelis (1990) stated the game-theoretic paradox that tighter sanctions need not improve behaviour. The selection insight—that observed sanctions are the failures, because threats that would work, work as threats—runs through Smith (1995), Drezner (2003), and Hovi et al. (2005), with Drezner (1999) the book-length statement and Lacy and Niou (2004) the issue-linkage variant; Morgan and Schwebach (1997) and Bapat and Kwon (2015) give the empirical and enforcement-theoretic counterparts. Our bands of protection failure are cousins of that selection logic with the mechanism relocated: threats are absent not because they would succeed unspoken but because announcing them costs more than they would spare. Closest in spirit is Verdier (2009), for whom sanctions are revelation regimes—instruments whose function is to sort types. This paper's analysis adds the accounting of who bears the cost of the sorting: the son.

Punishment, and punishing the punisher. Finally, the model inverts a familiar architecture. In the economics of punishment from Becker (1968) onward, and in the experimental literature on costly and third-party punishment (Fehr and

Gächter, 2000; Fehr and Fischbacher, 2004), sanctions discipline a wrongdoer on behalf of a community. Here the object of discipline is the punisher himself: the mother is a third-party punisher of a punisher, and the question is whether such second-order enforcement can protect the first victim. The answer the model returns—that it protects him only through the mother's resolve for further punishments being provable—has not been studied in that literature.

3 The model

A father wishes to punish a son, and would ideally inflict an amount of punishment that depends on his temperament, which only he knows: a just father wants a just amount, a harsh father an excessive amount. A mother cannot stop him directly, but she can make him pay. She can impose an immediate penalty on him, and she can threaten a further penalty to come. Whether she would actually follow through on the threat depends on her own resolve, which only she knows.

The order of events is as follows. The son has conducted his misdeed. That is in the past and irreversible. It is not part of the game I study. Rather, the game begins with the father's first strike. The mother, having observed the inflicted punishment, responds: she imposes a penalty now and announces a penalty scheme for what comes later. The father, having observed her response, strikes a second time and total punishment is then realised, as we restrict our analysis to two periods. Next, the mother decides whether to carry out the threatened penalty or not.

This is a game with two-sided incomplete information: initially, the mother does not know the father's type, nor does the father the mother's. But as the game unfolds, both will update, and in some cases the updating will collapse to certainty.

Figure 1 summarises the game's temporal structure.

Players and types. There are a father F, a mother M, and a son. The son is passive: he takes no action and is the object of the punishment; only F and M choose. Nature draws the two types independently,

$$i \in \{J, H\}, \quad \Pr(i = H) = \pi \in (0,1), \qquad j \in \{S, W\}, \quad \Pr(j = S) = \rho \in (0,1),$$

where i indexes the father's type and j the mother's. Each player observes only

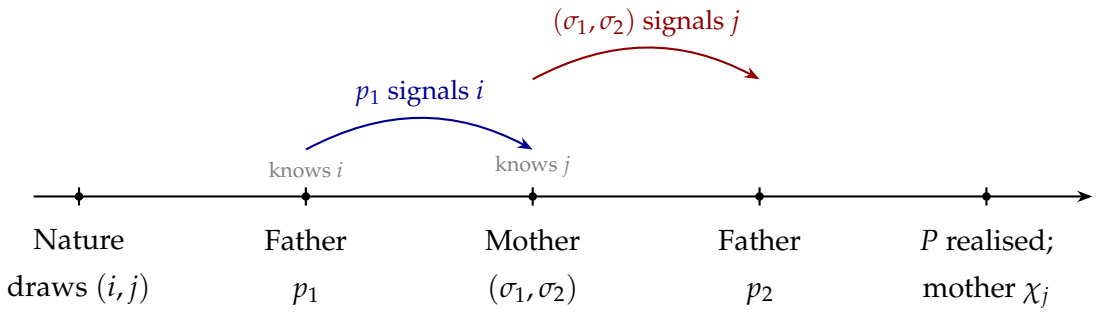

Figure 1: The order of play and the two signalling channels: the father's period-one punishment signals his type to the mother, and the mother's response signals her type to the father. At the final node $\chi_j \in \{0, 1\}$ is the mother's carry-out decision.

its own type.

Punishment. Punishment is delivered over two periods. In period $\tau \in \{1, 2\}$ the father inflicts $p_\tau \in [0, \bar{p}]$, where $\bar{p}$ is the per-period capacity; total punishment is $P = p_1 + p_2$. The father's ideal total is $y_J = 1$ for the just type and $y_H = \theta > 1$ for the harsh type.

There is no ambiguity among the players about what a just punishment entails—the mother agrees with the just father—which allows us to make the normalisation to 1.

The mother's instruments. After observing her own type and p_1, the mother picks two numbers: a period-one sanction σ_1 imposed immediately, and an *announced rate* σ_2 for period two, with $\sigma_1 \geq 0$ and $\sigma_2 \geq 0$: neither instrument is bounded above. The rate scheme threatens the father with a sanction $\sigma_2 (P - 1)$ proportional to the excess of the son's punishment over the just level, levied if the father pushes past $P = 1$ and the mother carries the scheme out. As a strategy these are rules $\sigma_1(p_1; j)$ and $\sigma_2(p_1; j)$ mapping the observed p_1 into the two numbers; at the point of choice they are simply scalars. At the final stage, after P is realized, she chooses whether to execute the announced scheme, $\chi_j \in \{0, 1\}$. We say *announced* advisedly: the announcement binds nothing by itself. Whether the scheme is carried out is the mother's final-stage choice.

Sanctions and costs. The realized sanctions are

$$\Sigma_1 = \sigma_1, \qquad \Sigma_2 = \chi_j \, \sigma_2 \, (P - 1) \, \mathbf{1}\{P > 1\} :$$

the period-one sanction is imposed in full by both types, while the period-two sanction, proportional to the excess $P - 1$, is realized only if the mother carries it out. The mother bears two linear costs. A *posturing cost* $\gamma\sigma_2$, with $\gamma > 0$, per unit of the *announced* rate σ_2, is paid by both types in period one upon announcing, whether or not the scheme is ever executed; it attaches only to the announcement, not to the act of carrying out. A *type-dependent cost* at rate c_j, with $c_S = 0$ and $c_W > 0$, applies to any sanction *actually imposed*—to Σ_1 and, when carried out, to Σ_2.

It may help to carry a gloss from the start. At the final stage the strong mother will behave exactly as a commitment type would: she executes what she announced, because nothing stops her—$c_S = 0$, and a convention of resolve introduced below closes the residual indifference. But she is not merely a commitment type, and the difference is the paper's hinge: her advantage is already present in period one, where the immediate sanction is free to her and dear to the weak type, and it is there—before any excess has occurred—that resolve can be proven. Section 6 shows that a model in which resolve manifests only at the moment of execution delivers none of what follows.

Payoffs. The father of type i receives

$$U_i = -\beta(P - y_i)^2 - \Sigma_1 - \Sigma_2, \qquad \beta > 0, \quad i \in \{J, H\},$$

where the scalar $\beta > 0$, common to both types, is the intensity of the punishment motive: it weights the deviation from the ideal y_i against the sanctions, which are denominated in a common currency. The mother of type j receives

$$U_j = -\ell(P) - c_j\Sigma_1 - \gamma\sigma_2 - c_j\Sigma_2, \qquad j \in \{S, W\},$$

with the quadratic loss

$$\ell(P) = (P - 1)^2.$$

The mother's ideal mimics the just father's ideal. She, too, wants just punishment to be carried out.

Strategies and beliefs. The father's period-one strategy is a number $p_1(i) \in [0, \bar{p}]$ for each type; his period-two strategy is a map $p_2(i, p_1, \sigma_1, \sigma_2) \in [0, \bar{p}]$, a function of his type, his own period-one choice, and the observed sanction pair.

The mother's strategy is the pair of rules $\sigma_1(p_1; j), \sigma_2(p_1; j)$ together with the carry-out χ_j. Strictly, this should also be a mapping from the game's full history

onto a binary decision but, as we will see, our assumptions about costs (and an additional tie-breaker assumption below) allow us to conflate it, considerably simplifying notation.

Two posteriors, one for each signalling channel, complete the description. After observing p_1, the mother updates on the father's type,

$$\hat{\pi}(p_1) = \Pr\big(i = H \mid p_1\big),$$

read against the father's period-one strategy $p_1(\cdot)$; before choosing p_2, the father updates on the mother's type,

$$\hat{\rho}\big(\sigma_1, \sigma_2; p_1\big) = \Pr\big(j = S \mid \sigma_1, \sigma_2, p_1\big),$$

read against the mother's rules at the realized p_1. The conditioning on p_1 is essential: the same sanction pair carries different information after a small p_1 than after a large one, because its likelihood is computed against $\sigma_1(p_1; j), \sigma_2(p_1; j)$, which themselves depend on p_1.

Two restrictions on parameters are maintained throughout.

Assumption 1. $\frac{1}{2} < \bar{p} < 1$ and $\theta \leq 2\bar{p}$.

Assumption 2. $\beta\gamma < \theta - 1$.

Assumption 1 ensures that neither ideal can be attained in a single period and that the harsh ideal is attainable in two. Assumption 2 gives deterrence a chance at all: it will emerge below that when $\beta\gamma \geq \theta - 1$, no mother of any type and any credibility ever finds any sanction worth its posturing, and the analysis collapses to the trivial regime in which a harsh father simply takes his ideal. (The object behind this threshold is the *target excess* b introduced in Section 4.4; the reader meeting the assumption here may take it on trust until then.) We keep that case in view as a limiting frame but exclude it from the formal treatment.

4 Analysis

4.1 Solution concept

We solve the game by backward induction for its perfect Bayesian equilibria: a profile of strategies for the father and the mother together with the posteriors $\hat{\pi}(p_1)$ and $\hat{\rho}(\sigma_1, \sigma_2; p_1)$, such that every action is sequentially rational given beliefs

and continuation play, and beliefs are derived from strategies by Bayes' rule wherever possible. We restrict attention to pure strategies, and we require equilibria to survive the Intuitive Criterion of Cho and Kreps (1987), applied to the mother's signalling problem at each realized p_1.

The game leaves behaviour undetermined at several points—ties, knife edges, a multiplicity of beliefs—and rather than resolve them piecemeal, we gather every selection rule into one place and use it consistently. All results below are statements about the equilibrium this block selects—a characterisation of the selected equilibrium, not of the full set of perfect Bayesian equilibria.

Definition 1 (The selected equilibrium)**.** Among the pure-strategy perfect Bayesian equilibria surviving the Intuitive Criterion, we select lexicographically by three conventions.

(R1) *Resolve.* The strong mother is lexicographically resolute: whenever she is otherwise indifferent at the final stage and the threatened sanction is positive, she carries out the announced sanction. This is where credibility enters the model, and it enters as a convention of resolve, not as a derived incentive (Remark 2).

(R2) *Parsimony.* No player takes a costly or idle action that serves no purpose: no mother imposes a sanction that neither disciplines nor informs; no father arms beyond what his ideal requires; the just father opens at the least reach that permits $P = 1$; and among signalling configurations that leave the sender's continuation payoff unchanged, the least signal is selected. In particular, immediate sanctions above the least separating level are discarded, and a weak mother exactly indifferent between revealing and mimicking the strong type's costly signal reveals.

(R3) *Quiet-side boundary defaults.* Knife-edge thresholds are closed on the quiet side: a strong mother exactly indifferent at the reticence-band edge—where the harm the threat would spare just equals its posturing cost—announces nothing, and a harsh father exactly indifferent between parking at the band edge and striking into separation parks. Given the just father's least opening, reaches above the just level are read as harsh: $\hat{\pi}(x) = 1$ for all $x > 1$.

Remark 1 (What the selection does, and what the refinement does). Three things of different rank happen in this block, and they should not be allowed to blur. The

first is the refinement, which is no convention at all: the Intuitive Criterion does substantive work in this model. It is what removes every pooling configuration (Lemma 9 below), and the removal of the pools is what exposes the weak mother—without it there is no unmasking, and no theorem. The second is (R1), resolve, the one convention with substance: it tilts the selection toward enforcement. Everything else is the resolution of ties on sets of measure zero, and it comes in two kinds. Parsimony strips waste—idle and over-sized sanctions, arms beyond the ideal—places the just father at his least sufficient opening, and resolves the weak mother's tie at mimicry indifference toward revelation. That last resolution leans against enforcement in form only: a weak type playing the strong bundle turns the configuration into a pool, Bayes' rule at the bundle diluting the very credence the mimicry was to exploit, and every such pool falls to the refinement (Lemma 9); the pretence, played, unravels what it borrows. The boundary defaults close knife edges of measure zero on the quiet side—the reticence band keeps its upper edge, parking is selected at first-strike indifference—which is what lets the father's problem attain its optimum: he will be seen to probe the very boundary of what the mother tolerates, and that boundary must be his to stand on. And the belief $\hat{\pi}(x) = 1$ extends the least opening: with the just father at reach 1, every reach above the just level is vacated for the harsh type alone. This is the belief most favourable to deterrence—just-father pooling at higher reaches would inflate the target excess $b = \beta\gamma/\hat{\pi}$ and *widen* the band of unanswered aggression, so admitting it only deepens the failures below. The theorems of this paper are therefore statements about enforcement *at its selected best*: where restraint breaks down here, the breakdown cannot be an artifact of how indifference was resolved.

4.2 The final stage

We begin at the last node, where the mother decides whether to carry out the announced sanction.

Lemma 1. *At any final history with a positive threatened sanction, $\sigma_2(P-1)\,\mathbf{1}\{P > 1\} > 0$, the weak mother strictly declines to carry it out and the strong mother is indifferent. At histories where the threatened sanction is zero, the carry-out decision is vacuous and parsimony selects no execution for both types.*

Proof. Immediate: all other terms are sunk, and executing costs the weak type $c_W\sigma_2(P-1) > 0$ and the strong type nothing, since $c_S = 0$. A zero threatened

sanction makes χ_j payoff-irrelevant for both types, an idle action removed by (R2). □

Remark 2 (The knife edge of resolve). By (R1) the strong mother carries out whenever there is anything to carry out. Her indifference is exact, since $c_S = 0$, and deserves a word rather than concealment. One reading is definitional: the strong type *is* the type for whom keeping her word costs nothing—resolve modelled as the absence of any reason to flinch. The other is a perturbation: an arbitrarily small expressive or reputational payoff to execution for the strong type breaks the indifference in favour of carrying out and changes nothing else. The same parameter makes the period-one sanction costless for her, so the benchmark throughout is one in which *proof is costly only for the wrong type*—the cleanest case for the signalling logic. A small positive cost for the strong type would survive intact wherever her stake in deterrence exceeds it, at the price of carrying the term through every formula.

4.3 The father's second strike

The just father. At the fourth stage the father chooses $p_2 \in [0, \bar{p}]$, adding to the sunk p_1, so the reachable totals form the interval $[p_1,\ p_1 + \bar{p}]$; since $\bar{p} < 1$ we have $p_1 < 1$ throughout. The just father wants $P = 1$, and as $p_1 < 1$ the only question is whether he has the capacity to reach it.

Lemma 2. *The just father sets $p_2 = 1 - p_1$ if $p_1 + \bar{p} \geq 1$, attaining $P = 1$; otherwise he sets $p_2 = \bar{p}$, attaining $P = p_1 + \bar{p} < 1$. In either case $P \leq 1$, so he never enters the sanctioned region, and the period-two sanction is irrelevant to his choice.*

Proof. His loss $\beta(P-1)^2$ is increasing in the distance of P from 1, and the reachable set is $[p_1, p_1 + \bar{p}]$ with $p_1 < 1$. The closest point to 1 is 1 itself when $1 \leq p_1 + \bar{p}$, and $p_1 + \bar{p}$ otherwise. Neither exceeds 1, so the sanction term never binds. □

The harsh father. The harsh father wants $P = \theta > 1$, so unlike the just father he wishes to push past the just level. What restrains him is the *expected* marginal sanction he faces above $P = 1$. He has observed the announced rate σ_2, but it bites only if the mother carries the scheme out; assigning probability $\hat{\rho}$ to her being strong, he expects a marginal sanction $\hat{\rho}\sigma_2$ on each unit of excess. He therefore chooses P to maximise $-\beta(P-\theta)^2 - \hat{\rho}\sigma_2(P-1)\,\mathbf{1}\{P > 1\}$ over the reachable

interval, whose top is

$$\bar{P} = \min\{p_1 + \bar{p},\ \theta\}.$$

The objective is concave, with unconstrained maximiser

$$P^{\circ} = \theta - \frac{\hat{\rho}\sigma_2}{2\beta}.$$

Write $r^{\star} = 2\beta(\theta - 1)$ for the *fully-deterring rate*: the expected marginal sanction at which the father is pulled all the way back to the just level.

Lemma 3. *The harsh father's second strike attains*

$$P^{\star} = \min\left\{\bar{P},\ \max\left\{1,\ \theta - \tfrac{\hat{\rho}\sigma_2}{2\beta}\right\}\right\}.$$

Equivalently, when $\bar{P} > 1$: he is held at $P = 1$ when $\hat{\rho}\sigma_2 \geq r^{\star}$ (full deterrence); he settles at the interior $P^{\circ} = \theta - \hat{\rho}\sigma_2/2\beta$ when $0 < \hat{\rho}\sigma_2 < r^{\star}$ and $P^{\circ} \leq \bar{P}$ (partial deterrence); and he is capped at the ceiling $\bar{P}$ otherwise, reaching his ideal θ when $p_1 + \bar{p} \geq \theta$. When $\bar{P} \leq 1$, the sanctioned region is unreachable and he chooses $P = \bar{P}$.

Proof. If $\bar{P} \leq 1$, the sanctioned region is unreachable, the scheme never bites, and the harsh father takes the largest feasible total, $P = \bar{P}$; the displayed formula returns exactly this, the inner maximum being at least 1. Let then $\bar{P} > 1$. Among totals below 1 no sanction is paid and the objective is strictly increasing toward $\theta > 1$, so any optimum lies in $[1, \bar{P}]$. On this interval the objective is concave with unconstrained maximiser P°, found from the first-order condition $2\beta(\theta - P) = \hat{\rho}\sigma_2$. Projecting P° onto $[1, \bar{P}]$ gives the stated $P^{\star}$: it equals 1 when $P^{\circ} \leq 1$, i.e. $\hat{\rho}\sigma_2 \geq 2\beta(\theta - 1) = r^{\star}$; it equals P° when P° lies inside; and it equals $\bar{P}$ when $P^{\circ} \geq \bar{P}$, the sanction then slack against the binding reach. Whenever $\bar{P} > 1$ the father never chooses below 1, which he prefers to any lower punishment; when $\bar{P} \leq 1$, capacity forces $P = \bar{P}$. □

Corollary 1. *To deter fully the announced rate must satisfy $\hat{\rho}\sigma_2 \geq r^{\star}$, that is $\sigma_2 \geq r^{\star}/\hat{\rho}$. The father's inference fixes the discount: a mother known weak ($\hat{\rho} = 0$) is never believed and never deters; a mother known strong ($\hat{\rho} = 1$) deters at the rate $r^{\star}$; under pooling ($\hat{\rho} = \rho$) the announced rate must be inflated to $r^{\star}/\rho$, since only the strong type would carry it out.*

4.4 The mother's problem at a given reach

In the third stage the mother makes two choices: the sanction σ_1 imposed directly, and the rate σ_2 she announces for period two, carried out if $P > 1$. One case can be set aside at once.

Lemma 4. *If $p_1 + \bar{p} \leq 1$, no deterrence is possible: even a harsh father cannot reach past the just level, so both types of mother set $\sigma_1 = \sigma_2 = 0$, and the son's punishment is at most $p_1 + \bar{p}$ regardless of either type.*

Proof. The reachable totals form $[p_1, p_1 + \bar{p}]$ with top at or below 1, so $P > 1$ is impossible and the scheme can never be triggered. No sanction affects the outcome; by parsimony neither type imposes one. The father, of either type, goes to the top $p_1 + \bar{p}$. □

We therefore assume $p_1 + \bar{p} > 1$ throughout the remainder. It is convenient to carry the analysis in terms of the father's *reach*

$$x \;\equiv\; p_1 + \bar{p},$$

the largest total he can still attain; since striking to reach past θ changes nothing—a harsh father never goes beyond his ideal—we may restrict attention to $x \in (1, \theta]$ without loss. Throughout this subsection the reach x and the mother's posterior $\hat{\pi} = \hat{\pi}(p_1)$ are fixed; both are endogenised in Section 5.

Notation. Write $a = \theta - 1$ for the harsh father's excess ideal. When the father attaches credibility $\hat{\rho}$ to the scheme, a mother weighing harm against posturing faces, per unit of announced rate, a pull-back of $\hat{\rho}/2\beta$ on the father and a cost of γ; equating the marginal harm reduction $\hat{\pi}\hat{\rho}(P-1)/\beta$ to γ gives the *target excess*

$$b(\hat{\rho}) \;=\; \frac{\beta\gamma}{\hat{\pi}\hat{\rho}},$$

the distance above the just level at which deterrence is no longer worth pushing. Two values of the credibility will matter: full credibility, with $b \equiv b(1) = \beta\gamma/\hat{\pi}$, and pooled credibility, with $b_p \equiv b(\rho) = \beta\gamma/(\hat{\pi}\rho)$. Finally, for $z \in (0, a]$ define

$$\Delta(z) \;=\; \sqrt{z\,(2a - z)}\,,$$

an increasing function with $\Delta(z) > z$ on $(0, a)$ and $\Delta(a) = a$. Its role is the following: $1 + z$ is where a mother *pulls the father to* if she deters at credibility-adjusted target excess z, while $1 + \Delta(z)$ is the reach below which she *declines to*

deter at all. The wedge between the two—the target is nearer the just level than the threshold for bothering—will do much of the work below.

We consider the two instruments in turn, asking through which the mother's types can separate—for it is separation that exposes the weak mother and so shapes the son's fate.

Remark 3. The sanction σ_1 is costly—it lowers the father's payoff by σ_1 and the mother's by $c_j\sigma_1$—yet it is sunk before the father's second strike, so it can bear on what follows only through the father's posterior. A positive σ_1 is therefore incurred only to signal—burned money, in the language of Section 2—and since it is free to the strong type but costs the weak type $c_W\sigma_1 > 0$, it is well suited to separating the types.

Lemma 5. *The types cannot separate through σ_2 alone.*

Proof. An announced rate σ_2 pulls the harsh father back continuously, by $\hat{\rho}\sigma_2/2\beta$, and costs the announcing mother the posturing $\gamma\sigma_2$; both the effect and the cost are the same whichever type announces it. So whatever rate the strong type announces, the weak type obtains the identical effect at the identical cost by matching it—and saving her son is worth that cost to her whenever it is worth it to the strong type. The rate σ_2 thus carries no differential that could make the types' choices diverge: any rate the strong type might announce to separate is one the weak type strictly wishes to match. □

Lemma 6. *The weak mother can reduce her son's punishment only by pooling with the strong type on a positive rate $\sigma_2 > 0$.*

Proof. Whenever the weak mother's type is revealed, the father learns she will not carry the scheme out, so he faces expected rate $\hat{\rho}\sigma_2 = 0$ and drives the son to the reach x, regardless of any rate she has announced. Only if she pools with the strong type does the father retain a positive belief $\hat{\rho} = \rho$; then an announced rate $\sigma_2 > 0$ leaves him facing expected rate $\rho\sigma_2 > 0$ and pulls him back below x. A pool on $\sigma_2 = 0$ secures no pull-back, so only pooling on a positive rate helps her. □

The strong mother believed. We look first at the benchmark in which the strong mother's type has been revealed, so the father holds $\hat{\rho} = 1$ and the weak type drops out of his reckoning. She still does not know the father's type, facing a

harsh one with probability $\hat{\pi}$, and only a harsh father ever triggers the scheme. Her problem is to choose the announced rate σ_2 to minimise $\hat{\pi}\big(P(\sigma_2)-1\big)^2+\gamma\sigma_2$, where by Lemma 3 a harsh father responds with $P(\sigma_2)=\min\{x,\ \max\{1,\ \theta-\sigma_2/2\beta\}\}$.

Lemma 7 (Reticence and deterrence)**.** *Let $\hat{\rho}=1$ and $b=\beta\gamma/\hat{\pi}<a$. The strong mother's announced rate takes one of two values:*

(i) *if $x\le 1+\Delta(b)$, she announces nothing, and a harsh father takes the reach x;*

(ii) *if $x>1+\Delta(b)$, she announces*

$$\sigma_2^\star \;=\; 2\beta\,(a-b),$$

holding a harsh father at $P=1+b$.

If $b\ge a$ she announces nothing at any reach. We call the interval $\big(1+b,\,1+\Delta(b)\big]$ the reticence band*: reaches at which she could profitably pull the father back, yet declines, because the harm spared does not cover the posturing.*

Proof. Rates below $2\beta(\theta-x)$ leave the father at the reach—the scheme is announced but never bites—so on $[0,2\beta(\theta-x)]$ the objective is $\hat{\pi}(x-1)^2+\gamma\sigma_2$, increasing, and the best point of the inactive range is $\sigma_2=0$. Rates $\sigma_2\ge 2\beta a$ fully deter, all inducing $P=1$, and are dominated by the lowest of them: same harm, higher posturing. On the active range $\big(2\beta(\theta-x),\,2\beta a\big]$ the objective is $\hat{\pi}(\theta-1-\sigma_2/2\beta)^2+\gamma\sigma_2$, convex, with first-order condition $\hat{\pi}(P-1)/\beta=\gamma$, i.e. $P=1+b$, attained at $\sigma_2^\star=2\beta(a-b)$; this lies in the active range precisely when $1+b<x$. (Feasibility of the target: the required second strike is $p_2=1+b-p_1$; from $1+b<x=p_1+\bar{p}$ it satisfies $p_2<\bar{p}$, and $p_2\ge 0$, i.e. $1+b\ge p_1$, is equivalent to $x-1-b\le\bar{p}$, which follows from $x\le\theta=1+a$, $b>0$, and $a\le 2\bar{p}-1<\bar{p}$.) The choice is therefore binary: announce nothing, for $\hat{\pi}(x-1)^2$, or announce $\sigma_2^\star$, for $\hat{\pi}b^2+\gamma\sigma_2^\star=\hat{\pi}b^2+2\hat{\pi}b(a-b)=\hat{\pi}\,b(2a-b)=\hat{\pi}\,\Delta(b)^2$. The strong mother thus weakly prefers deterrence iff $(x-1)^2\ge\Delta(b)^2$, i.e. $x\ge 1+\Delta(b)$; at equality, by (R3), the selected action is nevertheless silence ($\sigma_2=0$). Hence the selected action is silence for $x\le 1+\Delta(b)$ and deterrence strictly above. If $b\ge a$ the active-range objective is increasing throughout and she announces nothing. □

Two features deserve emphasis. First, when she deters, she stops short: even a

strong mother, certain of a harsh father and believed without doubt, leaves her son at $1 + b$, a residual harm above the just level that shrinks only as posturing becomes free. Second, *before* she deters there is a band of reaches in which she does nothing at all. Within the reticence band the father faces no sanction of any kind, from any mother, however she is believed. The band widens with $\beta\gamma$, and as $b \to a$ its upper edge $1 + \Delta(b) \to \theta$: reticence swallows deterrence continuously as threatening grows dear.

A word on the band's edge, which is where the quiet-side default (R3) does its only work. Strictly inside the band the strong mother declines to deter as a strict inequality—the harm spared falls short of the posturing, and her low cost of *carrying out* is beside the point, since what does not pay is the *announcement*. Strictly above it she deters, again strictly. Only at the single reach $x = 1 + \Delta(b)$ are the two exactly tied, and (R3) breaks that measure-zero tie toward silence. It does so not to restrain her—she is indifferent, and loses nothing—but to keep the father's problem well posed: he will want to park as high in the band as he can, and were the tie broken the other way his optimum would be a supremum he could approach but never attain. The convention buys existence of the parking outcome, nothing more.

Separation. Separation must run through σ_1 (Lemma 5), and the question is how large an immediate sanction the strong type must impose for the weak not to follow. Define the *mimicry stake*

$$M(x) \;=\; \hat{\pi}\Big[(x-1)^2 - \Delta(b)^2\Big]\,:$$

as the proof below shows, $M(x)$ is exactly what the weak type gains, gross of that sanction, by passing as strong rather than standing revealed.

Lemma 8 (Separation)**.** *Let $b < a$ and $x > 1 + \Delta(b)$. A separating equilibrium surviving the Intuitive Criterion exists at every such reach. In the selected one the strong type plays $(\sigma_1^{\text{sep}}, \sigma_2^{\star})$ with*

$$\sigma_1^{\text{sep}} \;=\; \frac{M(x)}{c_W},$$

the weak type plays $(0, 0)$, a harsh father proceeds to the reach x against her, and is held at $1 + b$ by a revealed strong mother.

Proof. Define the *value of credibility* at announced rate r: the harm a mother read

as strong is spared, net of the posturing,

$$H_x(r) \;=\; \hat{\pi}\big[(x-1)^2 - (P_x(r)-1)^2\big] - \gamma r.$$

On the dead range $r \le 2\beta(\theta - x)$ the response is $P_x = x$ and $H_x(r) = -\gamma r \le 0$; on the saturated range $r > 2\beta a$ the pull is capped at $h = 0$ and H_x falls in r, dominated by the active endpoint. On the active range $r \in \big(2\beta(\theta - x),\, 2\beta a\big]$, write $h = P_x(r) - 1$, so $r = 2\beta(a - h)$ and $\gamma r = 2\hat{\pi}b(a - h)$; then $H_x = \hat{\pi}[(x - 1)^2 - h^2 - 2b(a - h)]$, maximised at $h = b$, where

$$\max_r H_x(r) \;=\; \hat{\pi}\big[(x-1)^2 - (2ab - b^2)\big] \;=\; \hat{\pi}\big[(x-1)^2 - \Delta(b)^2\big] \;=\; M(x),$$

attained uniquely at $r = \sigma_2^\star$. So $M(x)$ is exactly the most that *being read as strong* can ever be worth, to either type, and the strong type's bundle attains it—the least-cost separating outcome in the sense of Riley (1979).

Three jobs remain, and it pays to keep them apart: incentive compatibility pins the immediate sanction from below; selection pins it from above and breaks a tie; the refinement then only checks beliefs off the path.

Incentive compatibility: the sanction from below. Revealed, the weak type announces nothing—any rate is discounted to zero and the posturing is pure waste—for a loss of $\hat{\pi}(x-1)^2$. Mimicking a bundle (σ_1, r) read as strong costs her the posturing and the immediate sanction but never the execution, so her gain over revelation is $H_x(r) - c_W\sigma_1 \le M(x) - c_W\sigma_1$. Any separating sanction paired with $\sigma_2^\star$ must therefore satisfy $c_W\sigma_1 \ge M(x)$: below that level she strictly mimics, which is why separation requires an immediate sanction at all.

Selection: the sanction from above, and the tie. At $\sigma_1 = M(x)/c_W$ the weak type is exactly indifferent between revealing and mimicking; revealing is then already a best response, and it is parsimony (R2), not the refinement, that selects it—the two replies leave her continuation payoff unchanged, the least signal is the non-signal, and the Intuitive Criterion disciplines beliefs after off-path deviations and has no purchase on an on-path indifference. Nor could the tie be broken the other way: a weak type playing the bundle turns the configuration into a pool, which Lemma 9 below removes. Immediate sanctions strictly above $M(x)/c_W$ also separate: the strong type does not bear them, and beliefs reading any lesser sanction as weak sustain them as perfect Bayesian equilibria surviving the refinement. They are, however, payoff-equivalent for the sender and strictly dearer for the father, and

(R2) discards them for the least sufficient sanction. (A divinity-style refinement in the spirit of Banks and Sobel (1987) would eliminate the over-sized sanctions directly, a downward deviation profiting the strong type over a strictly larger set of beliefs than the weak; under the Intuitive Criterion alone the elimination is by selection.)

Refinement: survival of the Intuitive Criterion. What remains for the refinement is the off-path beliefs at the selected configuration. A deviation (σ_1', r') is equilibrium-dominated for the weak type iff $H_x(r') - c_W\sigma_1' < 0$; for such deviations the criterion requires the father to place no weight on the weak type, reading them as strong. The strong type's gain from any deviation read as strong is $H_x(r') - M(x) \leq 0$: she already enjoys full credibility at the cheapest sufficient bundle, so no deviation the criterion forces to be read favourably strictly profits her. Deviations not dominated for the weak type may carry adverse beliefs, under which they are strictly unattractive to both types. The selected configuration survives. □

Remark 4. The separating sanction vanishes exactly at the edge of the reticence band: $\sigma_1^{\text{sep}} \downarrow 0$ as $x \downarrow 1 + \Delta(b)$, and $M(x) > 0$ precisely where the strong mother deters at all. Separation and deterrence switch on together.

Lemma 9 (Ambiguity does not survive scrutiny)**.** *No pooling equilibrium survives the Intuitive Criterion at any reach $x > 1 + \Delta(b)$.*

Proof. Fix any candidate pool: both types post the same (σ_1^p, σ_2^p), the father retains $\hat{\rho} = \rho$ and responds with some level P^p. Since neither type bears an execution cost and the immediate sanction enters separably, the part of the pooled loss *net of that sanction* is identical across types,

$$\Lambda^p \;=\; \hat{\pi}\,(P^p - 1)^2 + \gamma\sigma_2^p,$$

the weak type paying $c_W\sigma_1^p$ on top. Compare Λ^p with the corresponding net loss of full credibility at the cheapest sufficient rate, $\hat{\pi}\Delta(b)^2$. Writing $b_p = \beta\gamma/(\hat{\pi}\rho) > b$, the best net loss *any* pool can deliver is $\hat{\pi}\min\{(x-1)^2, \Delta(b_p)^2\}$ when $b_p < a$—deterring at diluted credibility forces an inflated announcement for the same pull, and whether it beats idling depends on the reach—and simply $\hat{\pi}(x-1)^2$, idling, when $b_p \geq a$. Since Δ is strictly increasing and $x > 1 + \Delta(b)$, both candidates strictly exceed $\hat{\pi}\Delta(b)^2$, so in either case

$$G^p \;\equiv\; \Lambda^p - \hat{\pi}\Delta(b)^2 \;>\; 0\,:$$

both types would gain exactly G^p, gross of any change in the immediate sanction, from being read as strong at the bundle $(\sigma_1', \sigma_2^\star)$.

Now use that σ_1 is unbounded: choose $\sigma_1' > \sigma_1^p + G^p/c_W$. For the weak type the deviation's value, even under the most favourable belief, is $G^p - c_W(\sigma_1' - \sigma_1^p) < 0$: it is strictly equilibrium-dominated. For the strong type the sanction is free, so under the belief $\hat{\rho} = 1$ the deviation yields exactly $G^p > 0$. The Intuitive Criterion therefore requires the father to read the deviation as strong, the strong type strictly profits, and the pool fails. The argument covers the silent pool $(0, 0)$, where $\Lambda^p = \hat{\pi}(x-1)^2$ and $G^p = M(x) > 0$. □

Remark 5 (What carries the no-pooling result). The lemma leans on two structural features, and both should be read in the open: the immediate sanction σ_1 is unbounded, and it is free to the strong type, $c_S = 0$. With a bounded sanction $\sigma_1 \leq \bar{\sigma}_1$, the pool-breaking deviation exists only if $\bar{\sigma}_1 - \sigma_1^p > G^p/c_W$; where the bound bites, pools can survive, and with them the weak mother's shelter. With a small positive cost $0 < c_S < c_W$ for the strong type—bracketing its knock-on effect at the final stage, Remark 2—the deviation requires a sanction increment d with $G^p/c_W < d < G^p/c_S$, which exists precisely because $c_S < c_W$. The engine is thus not unboundedness alone but an unbounded signal with a strict cost advantage for the resolute type; $c_S = 0$ is its cleanest case, in which proof is always technologically available and costly only for the wrong type.

Proposition 1 (The continuation at a given reach)**.** *Fix $\hat{\pi}$ and a reach $x > 1$, and write $b = \beta\gamma/\hat{\pi}$. If $b \geq a$, the selected continuation is* quiet *at every reach: no sanction from any mother, a harsh father takes the reach. If $b < a$, the selected surviving continuation is unique: quiet for $x \leq 1 + \Delta(b)$, and* separation *for $x > 1 + \Delta(b)$, as in Lemma 8. In either case the weak mother's son is never sheltered: where resolve can always be proven, irresolution can always be detected.*

Proof. If $b \geq a$, deterrence pays at no reach and no credibility: even at full credibility the active-range objective is increasing (Lemma 7), being read as strong is worth nothing, and by parsimony no mother sanctions or announces; the father takes the reach. Let then $b < a$. Below the band no deterrence is worth its posturing at any credibility (Lemma 7, with Δ increasing), so by parsimony all are silent. Above it, Lemma 9 removes every pool and Lemma 8 provides the separating equilibrium; by Lemma 6 the weak mother then helps her son not at

all. □

5 The father's first strike

We now step back to the father's opening move, taking Proposition 1 as the continuation. Because $\bar{p} < 1$ and $\bar{p} < \theta$, neither ideal can be delivered in one round: the first strike commits the father's reach $x = p_1 + \bar{p}$.

The just father. The just father attains his ideal at every opening whose reach lies in $[1, 2\bar{p}]$: there he can set $P = 1$, and by Lemma 2 he does. He does not open below reach 1, where his reach would fall short of the just level and he would undershoot $P = 1$ at a strict loss; and by parsimony (R2) he does not arm above what his ideal requires. He therefore opens at the least sufficient reach, reach 1. This is not innocuous, and the skeptical question deserves an answer in the open: could he not instead arm higher, pooling with the harsh type at some quiet-band reach and leaving the mother uncertain whom she faces? He could, at no cost to himself—but parsimony selects his least opening, and the choice is conservative in exactly the direction the paper cares about. Opening at reach 1 vacates every reach above the just level for the harsh type alone, which is what licenses the belief $\hat{\pi}(x) = 1$ of (R3); had the just father pooled higher, the posterior $\hat{\pi} < 1$ would only inflate $b = \beta\gamma/\hat{\pi}$ and widen the band, so every failure established below would deepen, not ease, under the alternative. No sanction is ever levied at his opening, the total being capped at 1 regardless of type (Lemma 4). For the remainder,

$$\hat{\pi} = 1, \qquad b = \beta\gamma, \qquad \Delta = \Delta(b), \qquad M(x) = (x-1)^2 - \Delta^2,$$

with $b < a$ by Assumption 2.

The value of a reach. The harsh father chooses $x \in [1, \theta]$, every value feasible by Assumption 1. His expected loss is

$$V(x) = \begin{cases} \beta(\theta - x)^2, & x \le 1 + \Delta, \\ L(x) \;=\; (1-\rho)\,\beta(\theta - x)^2 + \dfrac{\rho}{c_W}\,M(x) + \rho\,\beta\big(a^2 - b^2\big), & x > 1 + \Delta, \end{cases}$$

the separating branch collecting, as before, the immediate sanction $M(x)/c_W$ he absorbs, the quadratic loss at the hold level, and the executed sanction: $\beta(a - b)^2 + \sigma_2^\star b = \beta(a^2 - b^2)$.

Lemma 10. (i) *On the quiet band V is strictly decreasing.*

(ii) *Entering separation is discretely costly: at $x \downarrow 1+\Delta$, L exceeds the quiet value by $2\rho\beta a(\Delta - b) > 0$; the lower corner of the separating range is never optimal.*

(iii) *L is strictly convex with unconstrained minimiser*

$$x^\dagger = \frac{(1-\rho)\,\beta c_W\,\theta + \rho}{(1-\rho)\,\beta c_W + \rho} \;<\; \theta, \qquad x^\dagger - 1 = \frac{(1-\rho)\beta c_W}{(1-\rho)\beta c_W + \rho}\,(\theta - 1).$$

Proof. (i) Immediate. (ii) At $x = 1+\Delta$ the separating sanction vanishes (Remark 4) and the gap is $\rho\beta[(a^2 - b^2) - (a-\Delta)^2] = 2\rho\beta a(\Delta - b)$, using $\Delta^2 = 2ab - b^2$. (iii) The first-order condition $-2(1-\rho)\beta(\theta - x) + 2(\rho/c_W)(x-1) = 0$; the Δ^2 and constant terms drop out. □

Before stating the contest, pause on its shape, for the father's optimal reach will be discontinuous in the primitives, and the discontinuity has clean economics. Crossing the band edge is not a marginal act: the instant the reach exceeds $1+\Delta$, separation switches on and the father absorbs the expected sanction $\rho\, M(x)/c_W$ discretely—the entry toll of Lemma 10(ii). Marginal trespass is therefore dominated: whoever crosses at all must cross far enough for the interior optimum to repay the toll. The father either stands on the boundary or strikes deep; nothing in between is ever played. Figure 2 draws the two branches and the toll between them.

The father's problem is thus a contest between exactly two candidates: parking at the top of the quiet band, $x_A = 1+\Delta$, or striking at the interior optimum $x^\dagger$. Two conditions, both explicit in primitives, decide it. Write $m = (1-\rho)\beta$, $q = \rho/c_W$, and

$$\text{(P1)} \qquad (1-\rho)\,\beta c_W \;>\; \frac{\rho\,\Delta}{a-\Delta}, \tag{1}$$

$$\text{(P2)} \qquad \frac{mq}{m+q}\,a^2 \;-\; q\,\Delta^2 \;+\; \rho\beta\big(a^2 - b^2\big) \;<\; \beta\,(a-\Delta)^2. \tag{2}$$

(P1) is $x^\dagger > 1+\Delta$: the interior strike exists. (P2) is $L(x^\dagger) < V(x_A)$: striking beats parking, the left side being the closed form of $L(x^\dagger)$, its first term the harmonic-weighted minimum of the two quadratics.

Neither condition is free of γ, though the notation may hide it: $\Delta = \Delta(b)$ with $b = \beta\gamma$, so both (P1) and (P2) move with the posturing cost. As γ falls, Δ falls with it; the right side of (P1) collapses toward zero and (P2) relaxes alongside—which

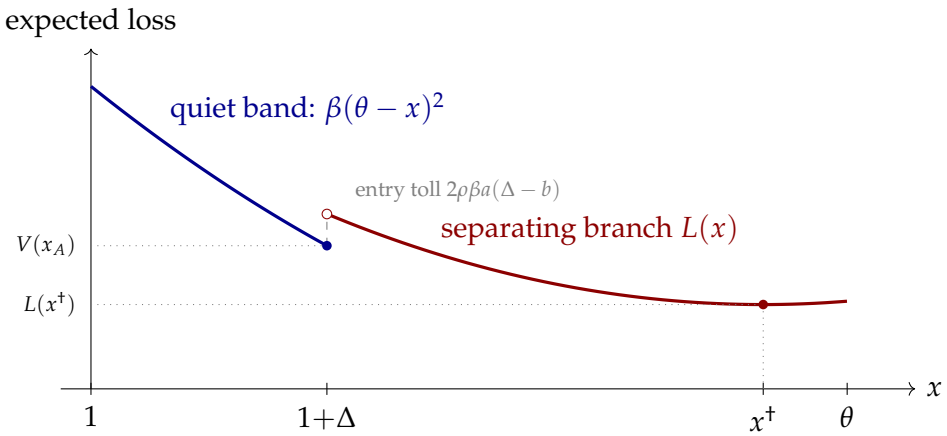

Figure 2: The harsh father's expected loss over the reach, drawn at illustrative parameters in the unmasking regime. On the quiet band the loss falls toward the band top $x_A = 1 + \Delta$; crossing the edge switches separation on, and the loss jumps by the entry toll of Lemma 10(ii) before descending, convexly, to the interior minimum $x^\dagger$. The contest is between the two marked candidates, x_A and $x^\dagger$; here (P1) and (P2) hold and $L(x^\dagger) < V(x_A)$, so the father strikes. Marginal trespass past the edge is dominated: whoever crosses must cross far.

is what powers the small-γ claim of the theorem below. Read in primitives, the unmasking regime is the land of cheap threats and dear follow-through: posturing inexpensive, so the band is thin and parking buys little; resolve rare and the weak mother's execution cost heavy, so the composite $(1-\rho)\beta c_W$ carries (P1). Where threatening is dear, the band swallows everything; where follow-through costs no one much, there is little to sort and the interior strike loses its edge.

Theorem 1 (The harsh father's strike, and the son's fate)**.** *Under Assumptions 1–2, in the selected pure-strategy perfect Bayesian equilibrium of Definition 1, for any $\gamma > 0$ the harsh father strictly arms: his reach satisfies $x^* \geq 1 + \Delta > 1$, so disarming—mimicking the just father—is never played. Moreover:*

(i) The unmasking (separation). *If* (P1) *and* (P2) *hold, he strikes $x^* = x^\dagger$; the mothers separate; the son is held at $1 + \beta\gamma$ when his mother is strong and driven to*

$$P_{\text{weak}} \;=\; x^\dagger \;=\; \theta - \frac{\rho\,(\theta - 1)}{(1-\rho)\,\beta c_W + \rho}$$

when she is weak.

(ii) The band (silence). *Otherwise he parks, $x^* = 1 + \Delta$; no sanction is ever levied, and the son's punishment is $1 + \Delta(\beta\gamma)$ under* both *mother types.*

Both conditions (P1) *and* (P2) *hold on an open set of primitives; each depends on* γ *through* $\Delta = \Delta(\beta\gamma)$*, and for any* β, ρ, c_W *they hold for all* γ *small enough.*

Proof. Reach 1 is the left endpoint of the quiet band, on which V is strictly decreasing and sanction-free, and $\Delta > 0$: he never disarms. By Lemma 10 the optimum is x_A or the projection of $x^\dagger$; the projection is interior iff (P1), in which case (P2), holding strictly, decides; at equality in (P2) both reaches are optimal and (R3) selects parking. When (P1) fails the corner is dominated by x_A via the jump. Outcomes are Proposition 1 and Lemma 8. Non-vacuousness: as $\gamma \to 0$, $b \to 0$ and $\Delta \to 0$; (P1) tends to $(1-\rho)\beta c_W > 0$, and (P2) tends to $\frac{mq}{m+q}a^2 + \rho\beta a^2 < \beta a^2$, which, dividing by a^2 and using $m = (1-\rho)\beta$, is $\frac{mq}{m+q} < m$, i.e. $\frac{q}{m+q} < 1$—both strict. □

Remark 6 (Sanity check at the extremes). As $\beta\gamma \to 0$ and $(1-\rho)\beta c_W/\rho \to \infty$ the fates sort completely: the strong mother's son is held within any ε of the just level, the weak mother's son driven within ε of θ. In the depth of the unmasking regime, deterrence does not fail on average—it *sorts*: the son's fate is decided entirely by his mother's resolve, and the father arms hard precisely because weakness will be found out.

Proposition 2 (Comparative statics)**.** *In the unmasking regime:*

(i) $P_{\text{weak}} = x^\dagger$ *is strictly increasing in* β*,* c_W *and* θ*, strictly decreasing in* ρ*, and independent of* γ*;*

(ii) *the strong-branch outcome* $1 + \beta\gamma$ *is increasing in* β *and* γ*;*

(iii) *the expected punishment level* $\rho(1+\beta\gamma) + (1-\rho)\,x^\dagger$ *is strictly decreasing in* ρ *and strictly increasing in* β*.*

In the band regime, the son's punishment $1 + \Delta(\beta\gamma)$ *is strictly increasing in* β *and* γ *under both mother types.*

Proof. (i) Write $A = (1-\rho)\beta c_W$, so $x^\dagger - 1 = a\,A/(A+\rho)$, increasing in A and in a. Explicitly,

$$\frac{\partial x^\dagger}{\partial \rho} = -\frac{a\,\beta c_W}{\big((1-\rho)\beta c_W + \rho\big)^2} < 0, \qquad \frac{\partial x^\dagger}{\partial \beta} = \frac{a\,(1-\rho)c_W\,\rho}{\big((1-\rho)\beta c_W + \rho\big)^2} > 0,$$

and symmetrically for c_W; monotonicity in θ is monotonicity in a. (ii) Immediate. (iii) In ρ the derivative is $(1+\beta\gamma) - x^\dagger + (1-\rho)\,\partial x^\dagger/\partial\rho$, and both pieces are

negative: the first because $x^\dagger > 1 + \Delta > 1 + b$ in the regime, the second by (i). In β both branches rise: the strong branch at rate $\rho\gamma > 0$, the weak branch at rate $(1-\rho)\,\partial x^\dagger/\partial\beta > 0$. The band statement is $\Delta'(z) = (a-z)/\Delta(z) > 0$. □

These are within-regime statements, and the caveat matters most for γ. The independence of $x^\dagger$ from γ holds only while the unmasking is selected: raising γ leaves $x^\dagger$ in place but lifts the band top $1 + \Delta(\beta\gamma)$ toward it, and once (P1) fails the equilibrium switches into the band—where γ is anything but neutral. The handover is taken up in Corollary 2.

One derivative in (i) deserves to be lifted from the list, for it is the least intuitive in the paper: P_{weak} *rises* in c_W. One might have guessed the opposite—the weak mother's burden, surely, is her own affair. It is not, and the chain runs through the father's pocket. A larger c_W makes mimicry dearer, which cheapens the strong mother's proof: the separating sanction $\sigma_1^{\text{sep}} = M(x)/c_W$ falls. That sanction is precisely what the father absorbs when the mother proves strong, and as it shrinks, so does his reason for restraint: he arms harder, and the reach he commits—the reach the *weak* mother's son will absorb in full—grows. The parameter that makes resolve provable is the one that drives the unprotected son's exposure; Section 6 will give the phenomenon its name, the father's information rent, collected from the boy.

Corollary 2 (Zeal)**.** *The father's intensity is malign through every channel that exists. Within the unmasking regime it raises both branches at once; within the band regime it raises the unanswered reach; and in the joint limit $\beta\gamma \to \theta - 1$ the son's punishment converges to the harsh ideal θ under* both *mother types and in* whichever *regime is selected.*

Proof. The first two claims are Proposition 2. For the limit, the selected outcome is $1 + \Delta$ in the band and $\{1 + b,\ x^\dagger\}$ in the unmasking, where (P1) places $x^\dagger$ strictly *above* the band top: $1 + \Delta < x^\dagger \le \theta$ whenever the unmasking is played. As $\beta\gamma \to a$, both $1 + b \to \theta$ and $1 + \Delta \to \theta$, and the sandwich $1 + \Delta < x^\dagger \le \theta$ forces $x^\dagger \to \theta$ along any sequence on which the unmasking is selected. Which regime survives near the limit depends on the path: $x^\dagger$ is γ-free, so raising γ at fixed (β, ρ, c_W) leaves it in place while the band top rises toward it, and (P2) fails before (P1) can bind: where (P1) binds, $x^\dagger = 1 + \Delta$, and the entry jump of Lemma 10(ii) gives $L(x^\dagger) = V(x_A) + 2\rho\beta a(\Delta - b) > V(x_A)$, so (P2) has already

failed—by continuity, strictly before. The band therefore takes over before the limit is reached, and the convergence claim is indifferent to the handover. □

Corollary 3. *Any path that offers the son any protection—any outcome in which the father is held short of his committed reach—necessitates a sanction in period one.*

Proof. By Proposition 1 the only continuation in which the father is held short of his reach is separation, which requires the positive immediate sanction $\sigma_1^{\text{sep}} = M(x)/c_W > 0$ (Lemma 8); the announced rate alone cannot separate (Lemma 5) and no pooled announcement survives (Lemma 9). Note the timing: $p_1 = x - \bar{p} \leq \bar{p} < 1$ by Assumption 1, so the sanction is levied while the punishment inflicted still lies below the just level. □

6 Two ways to tragedy

Theorem 1 leaves the harsh father exactly two configurations, and they are the two ways to tragedy.

The band: quiet harm. In the first (trivial) way to tragedy the harsh father parks at the top of the reticence band, $1 + \Delta(\beta\gamma)$, and *the mothers stay silent*: no immediate sanction, no announced sanctioning scheme, leaving the son—whoever his mother is—exposed to a punishment exceeding the just level. The mother is silent not because she is weak but because answering is not worth her posturing; the father, knowing the exact boundary of what will be tolerated, stands on it. As threatening grows dearer or the father more zealous, the band widens, and at $\beta\gamma \to \theta - 1$ the harsh father can reach his ideal.

The unmasking: sorted harm. In the second (less trivial) way to tragedy the harsh father exposes himself, screening the mothers and unmasking the weak one. The resolute mother proves herself and holds the son at $1 + \beta\gamma$; the weak mother's son, by contrast, absorbs the father's entire chosen reach. Fates sort, and the sorting is the tragedy: the son's punishment is decided not by what he did, nor even by the father alone, but by whether his mother can afford to be believed.

What restrains the father, and what does not. The harsh father's restraint is priced by a single object—the sanction $M(x)/c_W$ the father absorbs when the mother proves strong.

Three readings of the same formula $x^{\dagger} = \theta - \rho(\theta - 1)/[(1-\rho)\beta c_W + \rho]$ organise the comparative statics.

Credibility restrains before the fact: a larger ρ lowers the damage *even conditional on the mother proving weak*, through the reach the father declined while he still fears the presence of the strong mother.

Zeal dissolves restraint: as β grows the father ceases to weigh the sanction, and $x^{\dagger} \to \theta$. And the weak mother's pain is the father's information rent: a larger c_W cheapens the strong mother's proof, shrinks the sanction the father absorbs, and so *emboldens* him—the very parameter that makes resolve provable is the one that drives the unprotected son's exposure.

What resolve must be made of. The assumption $c_S = 0$ may look like a normalisation. It is not; it is load-bearing, and the cleanest way to see it is to solve the model it excludes. Suppose instead $c_S = c_W = c > 0$, and let the strong mother be a commitment type in the tradition of Kreps and Wilson (1982): an automaton who executes whatever she has announced. Resolve is then perfectly real—but it manifests only *ex post*, at the moment of execution, and the architecture collapses. The immediate sanction now costs both types alike, so it loses the differential that let it separate: the argument of Lemma 5 extends to it verbatim. Worse, the announcement channel runs backwards. The committed type expects to pay $c\,\sigma_2(P-1)$ upon execution; the weak type knows she will renege and bears only the posturing; announcing is now *dearer for the resolute type*—the wrong Spence ordering, mimicry cheap for exactly the type one wishes to screen out. And the deviation that broke the pools required, as Remark 5 recorded, an unbounded signal with a strict cost advantage for the resolute type; here there is none, and the pools stand. The full taxonomy of that model's equilibria is another paper's work, but this much is immediate: nothing in it separates, no mother is ever found out, and none is ever proven. The lesson is Corollary 3 restated at the level of model architecture. Resolve that shows itself only at execution protects no one, because it arrives after everything has been decided; protection requires resolve that is demonstrable in advance, through a cost advantage in the signalling stage—and $c_S = 0$ is the cleanest world in which such proof exists.

Protection must be paid for early. Every path on which the son is protected runs through the period-one sanction (Corollary 3). And the sanction comes early by necessity: it is paid while the punishment inflicted so far still lies below the just

level, answering not an excess—none has yet occurred—but an armament, the reach the father has committed. This is the model's most uncomfortable political lesson. Whoever wants to protect must sanction before the wrong, on the evidence of preparation alone; a third party that waits until the excess has materialised has nothing left to prove and no one left to deter. Even the father's restraint short of his ideal is priced by the same early instrument: what holds his reach down is the sanction he expects to absorb should the mother prove strong.

What the model finally says. Conditional on a harsh father, the son is never punished at the just level, and near it only when his mother is resolute and the father's strike gives her the occasion to prove it. Every other path is one of the two tragedies: the band, where harm is quiet and universal; the unmasking, where the weak mother's son gets exposed and suffers badly. Both ways to tragedy widen in the father's zeal; each also widens in its own cost—the band in the price of threatening, the unmasking in the weak mother's price of acting. The regions of greatest harm are intuitive: a zealous father, expensive threats, rare resolve, and a pretence too painful to maintain—there the model places the weak mother's boy at the edge of θ.

7 Coda

The formalism is done; what follows is interpretation, and the reader who wants only the theory may stop here.

Gaza. The model was written with Gaza in view, and its mapping is as plain as it is partial. A power inflicts punishment on a population beyond what even sympathetic observers would call just; states and communities of states weigh sanctions; some announce; few are believed; the punishment runs on. The model adds to this familiar picture one disciplined claim: that the outcomes we observe need not be evidence of indifference. The band is what non-enforcement looks like when announcing costs more than it spares; the unmasking is what enforcement looks like when proof is possible—and neither failure requires anyone to have wished the population ill. That is not a comfort. It is, if anything, the opposite: the tragedy survives good intentions, full information about the stakes, and unbounded capacity to sanction. And what the model asks of a protector is harder than any of these: to sanction before the excess, on the evidence of armament alone—for a sanction that waits until the wrong is complete protects no one

(Corollary 3).

Red lines, and due process. A reader of an earlier draft pointed the model at another episode: the American red line over chemical weapons in Syria—an announcement, prominent and public, with no sunk cost attached, and in the end no execution. From the vantage point of this paper's model this hardly comes as a surprise. Where an initial sanction would have been on the table, talk about future sanctions is not merely cheap—it is testimony against the speaker.

Yet the model's prescription—sanction before a verified escalation to excessive harm, on the evidence of some inflicted harm alone—collides with norms of due process. This is made worse by the fact that if early sanctions do their work and prevent excessive harm, the counterfactual is never observed and the protecting mother who acts early can never prove the necessity of what she did. This tension is not resolvable within the model, and I will not pretend otherwise. What the model contributes is the price of resolving it in due process's favour: a third party that waits until the wrong is complete has nothing left to prove and no one left to deter. The natural extension—a noisy signal of the reach, so that early sanctions carry error costs and legitimacy costs of their own—is where the dilemma would become tractable; here it is only named.

The weak mother. One sanctioner sits uneasily close to the model's weak type. Germany has declared Israel's security its *Staatsräson*, and the declaration has a precise translation in the model's parameters: it is c_W—the cost of carrying a sanction out, made immense by history. The model does not say such a mother is insincere; her announcements may be entirely earnest. It says she will be *found out*: in any equilibrium where proof circulates, the type that cannot afford execution is identified exactly because others can afford the proof she cannot match. And then the model's coldest line applies: the weak mother's pain is the father's information rent. The deeper the bind that makes her sanctions incredible, the cheaper it is for the world to be sorted into those who would act and those who would not—and the harder the father can afford to arm. Nothing in this reading requires bad faith anywhere. The rent is collected in equilibrium, and it is paid by the boy.

References

Abreu, D., and F. Gul (2000). Bargaining and reputation. *Econometrica* 68(1), 85–117.

Austen-Smith, D., and J. S. Banks (2000). Cheap talk and burned money. *Journal of Economic Theory* 91(1), 1–16.

Baliga, S., and T. Sjöström (2004). Arms races and negotiations. *Review of Economic Studies* 71(2), 351–369.

Banks, J. S., and J. Sobel (1987). Equilibrium selection in signaling games. *Econometrica* 55(3), 647–661.

Bapat, N. A., and B. R. Kwon (2015). When are sanctions effective? A bargaining and enforcement framework. *International Organization* 69(1), 131–162.

Becker, G. S. (1968). Crime and punishment: an economic approach. *Journal of Political Economy* 76(2), 169–217.

Ben-Porath, E., and E. Dekel (1992). Signaling future actions and the potential for sacrifice. *Journal of Economic Theory* 57(1), 36–51.

Cho, I.-K., and D. M. Kreps (1987). Signaling games and stable equilibria. *Quarterly Journal of Economics* 102(2), 179–221.

Crawford, V. P., and J. Sobel (1982). Strategic information transmission. *Econometrica* 50(6), 1431–1451.

Drezner, D. W. (1999). *The Sanctions Paradox: Economic Statecraft and International Relations*. Cambridge: Cambridge University Press.

Drezner, D. W. (2003). The hidden hand of economic coercion. *International Organization* 57(3), 643–659.

Eaton, J., and M. Engers (1992). Sanctions. *Journal of Political Economy* 100(5), 899–928.

Eaton, J., and M. Engers (1999). Sanctions: some simple analytics. *American Economic Review* 89(2), 409–414.

Fearon, J. D. (1994). Domestic political audiences and the escalation of international disputes. *American Political Science Review* 88(3), 577–592.

Fearon, J. D. (1995). Rationalist explanations for war. *International Organization* 49(3), 379–414.

Fearon, J. D. (1997). Signaling foreign policy interests: tying hands versus sinking costs. *Journal of Conflict Resolution* 41(1), 68–90.

Fehr, E., and U. Fischbacher (2004). Third-party punishment and social norms. *Evolution and Human Behavior* 25(2), 63–87.

Fehr, E., and S. Gächter (2000). Cooperation and punishment in public goods experiments. *American Economic Review* 90(4), 980–994.

Hovi, J., R. Huseby, and D. F. Sprinz (2005). When do (imposed) economic sanctions work? *World Politics* 57(4), 479–499.

Kaempfer, W. H., and A. D. Lowenberg (1988). The theory of international economic sanctions: a public choice approach. *American Economic Review* 78(4), 786–793.

Kartik, N. (2009). Strategic communication with lying costs. *Review of Economic Studies* 76(4), 1359–1395.

Kartik, N., M. Ottaviani, and F. Squintani (2007). Credulity, lies, and costly talk. *Journal of Economic Theory* 134(1), 93–116.

Kreps, D. M., and R. Wilson (1982). Reputation and imperfect information. *Journal of Economic Theory* 27(2), 253–279.

Kydd, A. H. (2005). *Trust and Mistrust in International Relations*. Princeton: Princeton University Press.

Lacy, D., and E. M. S. Niou (2004). A theory of economic sanctions and issue linkage: the roles of preferences, information, and threats. *Journal of Politics* 66(1), 25–42.

Milgrom, P., and J. Roberts (1982). Predation, reputation, and entry deterrence. *Journal of Economic Theory* 27(2), 280–312.

Morgan, T. C., and V. L. Schwebach (1997). Fools suffer gladly: the use of economic sanctions in international crises. *International Studies Quarterly* 41(1), 27–50.

Powell, R. (2006). War as a commitment problem. *International Organization* 60(1), 169–203.

Riley, J. G. (1979). Informational equilibrium. *Econometrica* 47(2), 331–359.

Sartori, A. E. (2002). The might of the pen: a reputational theory of communication in international disputes. *International Organization* 56(1), 121–149.

Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schelling, T. C. (1966). *Arms and Influence*. New Haven: Yale University Press.

Schultz, K. A. (1998). Domestic opposition and signaling in international crises. *American Political Science Review* 92(4), 829–844.

Sechser, T. S. (2010). Goliath's curse: coercive threats and asymmetric power. *International Organization* 64(4), 627–660.

Slantchev, B. L. (2005). Military coercion in interstate crises. *American Political Science Review* 99(4), 533–547.

Smith, A. (1995). The success and use of economic sanctions. *International Interactions* 21(3), 229–245.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics* 87(3), 355–374.

Tsebelis, G. (1990). Are sanctions effective? A game-theoretic analysis. *Journal of Conflict Resolution* 34(1), 3–28.

Verdier, D. (2009). Sanctions as revelation regimes. *Review of Economic Design* 13(3), 251–278.